TRIBUNAL DE CONTAS DA UNIÃO
INSTITUTO SERZEDELLO CORRÊA
ESCOLA SUPERIOR DO TRIBUNAL DE CONTAS DA UNIÃO

MARCUS VINÍCIUS BORELA DE CASTRO

Mineração de Dados com Rastro: Boas Práticas para Documentação de Processos e sua Aplicação em um Projeto de Classificação Textual

Brasília
2019

MARCUS VINÍCIUS BORELA DE CASTRO

Mineração de Dados com Rastro:
Boas Práticas para Documentação de Processos
e sua Aplicação em um Projeto de Classificação Textual

Trabalho de conclusão do curso de pós-graduação *lato sensu* em Análise de Dados para o Controle realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista em Análise de Dados.

Orientador: Prof. Dr. Remis Balaniuk

Brasília

2019

REFERÊNCIA BIBLIOGRÁFICA

CASTRO, Marcus Vinícius Borela. Mineração de Dados com Rastro: Boas Práticas para Documentação de Processos e sua Aplicação em um Projeto de Classificação Textual. 2019. Trabalho de Conclusão de Curso (Especialização em Análise de Dados para o Controle) – Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa, Brasília DF.

CESSÃO DE DIREITOS

NOME DO AUTOR: Marcus Vinícius Borela de Castro
TÍTULO: Mineração de Dados com Rastro: Boas Práticas para Documentação de Processos e sua Aplicação em um Projeto de Classificação Textual
GRAU/ANO: Especialista/2019

_____

Marcus Vinícius Borela de Castro
borela@tcu.gov.br

Ficha catalográfica

Castro, Marcus Vinícius Borela de

Mineração de dados com rastro: boas práticas para documentação de processos e sua aplicação em um projeto de classificação textual / Marcus Vinícius Borela de Castro; orientador, Remis Balaniuk, 2019.

109 p.

Monografia (especialização) - Escola Superior do Tribunal de Contas da União, Curso de Especialização em Análise de Dados para o Controle, Brasília, 2019.

Inclui referências.

1.Análise de Dados. 3. Mineração de Dados. 4. Classificação Textual. 5. Aprendizado de Máquina. 6. Boas práticas. I. Balaniuk, Remis. II. Escola Superior do Tribunal de Contas da União. Especialização em Análise de Dados para o Controle. III. Título.

Marcus Vinícius Borela de Castro

Mineração de Dados com Rastro:
Boas Práticas para Documentação de Processos
e sua Aplicação em um Projeto de Classificação Textual

Trabalho de conclusão do curso de pós-graduação *lato sensu* em Análise de Dados para o Controle realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista em Análise de Dados.

Brasília, 15 de agosto de 2019.

Banca Examinadora:

_____

Prof. Dr. Remis Balaniuk
Orientador
Instituto Serzedello Corrêa - TCU

_____

Prof. Dr. Edans Flávius de Oliveira Sandes
Instituto Serzedello Corrêa - TCU

À minha esposa, companheira no projeto da vida, Ana Cláudia, por ter me proporcionado minerar padrões eternos de amor nos registros da minha existência e me proporcionado um rastro permanente de felicidade na vida!

AGRADECIMENTOS

A gratidão é o principal passo em direção a um salto na felicidade. Gostaria de aproveitar a oportunidade para eternizar alguns agradecimentos a pessoas especiais.

Agradeço à minha esposa Ana Cláudia Lopes de Castro, que sempre esteve ao meu lado, nos bons momentos e nos difíceis, não só neste trabalho, mas por toda uma vida!

Aos meus filhos, Vinícius Lopes Borela de Castro e Bruno Lopes Borela de Castro, que me motivam diariamente a buscar o melhor em mim e que são os modelos com maior acurácia gerados no meu projeto de existência.

Agradeço aos meus amados e saudosos pais, José Lacerda de Castro, Zequinha, e Ilda Borela de Castro, origem de tudo, que sei, de onde estiverem, partilham dos meus projetos. Estendo esse agradecimento aos meus irmãos Marisa Borela de Castro Abelha e José Antônio Borela de Castro e todos seus filhos e netos.

Ao meu orientador Remis Balaniuk que aceitou o desafio de orientar o velho amigo, de 3 décadas, e alcançou elevada precisão em suas críticas construtivas e soube lidar bem com a ignorância dupla do amigo: ignorância no saber e ignorância na vontade de querer fazer a diferença.

Aos meus colegas de pós-graduação, que durante praticamente um ano estiveram ombreados a mim na missão da conquista de um complexo e desafiador conhecimento de Análise de Dados. Ressalto uma especial gratidão aos colegas Leonardo Augusto da Silva Pacheco e Renata Guanaes Machado, que não só estiveram sempre ao meu lado fisicamente nesse período, vizinhos à esquerda e à direita, respectivamente, como também enfrentaram comigo os diversos trabalhos das disciplinas do curso, laboratórios que promoveram não só o nosso crescimento intelectual mas estreitaram nossas amizades.

Aos professores do curso que, em sua grande maioria, planejaram com dedicação suas aulas e as atividades extras. Esse carinho nos alcançou a todos e foi fundamental para o nosso aprendizado. Destaco uma gratidão especial ao professor Edans Flavius de Oliveira Sandes que prontamente aceitou contribuir com esse trabalho ao compor a banca examinadora e ao professor Erick Muzart Fonseca dos Santos que incentivou a criação do Rastro-DM.

Ao ISC (Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa) e seus colaboradores. O ambiente de ensino esteve o tempo todo impecável, seja no clima amigável do espaço, seja no adequado suporte técnico às atividades de classe.

Aos colegas que integram a equipe de desenvolvimento do sistema e-TCE que me proporcionaram o primeiro desafio corporativo na área de mineração de dados, a construção do Cladop, um classificador textual para documentos do sistema. Destaco um agradecimento especial aos colegas André Augusto Siqueira, Larissa Beatriz de Souza Maia, minha querida comadre, e Rodrigo de Castro Soares, que acompanharam mais de perto o projeto.

Aos colegas de área do TCU, por serem exemplos de competência e de companheirismo. Destaco dois agradecimentos especiais: ao meu amigo Paulo Fonseca Merçon, amigo de algumas décadas, sempre um suporte de sabedoria e de humildade! E ao meu vizinho de mesa Luis André Dutra e Silva, vizinho também de terra natal, nasceu em Caratinga, próximo de Manhuaçu, MG. Autodidata, desbravador da área de *Machine Learning*, e detentor de um enorme coração: sempre apoia os colegas nos difíceis passos por esse desafiador caminho.

Não basta amar as pessoas, elas precisam se sentir amadas!

(Dom Bosco, século XIX)

RESUMO

Este trabalho propõe um conjunto de boas práticas de documentação de projetos de mineração de dados (DM), Rastro-DM, com foco não no modelo gerado, mas no processo por trás de sua construção, de forma a deixar um rastro das ações planejadas, dos treinamentos realizados, dos resultados obtidos e dos aprendizados concebidos. As práticas propostas são complementares às metodologias estruturantes de DM, tal como o CRISP-DM, que trazem todo o arcabouço metodológico e paradigmático para o processo de DM. Ilustra-se o seu uso em um projeto de classificação textual de documentos em PDF associados a danos ao Erário Público Federal Brasileiro denominado Cladop. Mostra-se, no contexto do Cladop, o uso do rastro documental para a geração semi-automática de relatórios e a sua integração com uma rotina de monitoramento automático proposta para classificadores em produção. A construção do kit Rastro-DM em um projeto é um pequeno passo que pode levar a um salto organizacional, a ser obtido com a partilha e o uso do rastro de forma corporativa.

Palavras-chave: Mineração de dados. Análise de dados. Ciência de dados. Aprendizado de máquina. Conhecimento organizacional. Metodologia. Boas práticas. Análise de dados no Governo. Classificação textual. Documentação. Documentação de projetos de mineração de dados. Geração automática de relatórios. Monitoramento de classificadores.

ABSTRACT

Data Mining with Trail: Best Practices for Process Documentation and its Application in a Textual Classification Project

This paper proposes a set of best practices for documentation of data mining (DM) projects, Rastro-DM, with a focus not on the model that is generated, but on the process behind its construction, in order to leave a trail of planned actions, trainings performed, results obtained and lessons learned. The proposed practices are complementary to the structuring methodologies of DM, such as CRISP-DM, which estabilish a methodological and paradigmatic framework for the DM process. The application of best practices is illustrated in a project called Cladop that was created for the classification of PDF documents associated with the investigatory process of damages to the Brazilian Federal Public Treasury. Two benefits are shown in the context of Cladop: the use of the document trail for semi-automatic report generation and its integration with an automatic monitoring procedure proposed for classifiers in production. Building the Rastro-DM kit in the context of a project is a small step that can lead to a institutional leap to be achieved by sharing and using the trail across the enterprise.

Keywords: Data mining. Data analysis. Data science. Machine learning. Organizational knowledge. Methodology. Best Practices. Data analysis in Government. Textual classification. Documentation. Documentation of data mining projects. Automatic Reporting. Monitoring procedure for classifiers.

SUMÁRIO

1 INTRODUÇÃO

Os dados estão mudando tudo e a capacidade de manipulá-los e entender Ciência de Dados está se tornando cada vez mais crítica para atuais e futuras descobertas e inovações (BERMAN *et al*, 2018).

Projetos de mineração de dados (DM) são desafiadores não só pelo complexo processo usado, exploratório, mas também por, em geral, serem inovadores, únicos e muitas vezes desenvolvidos por indivíduos ou pequenas equipes.

Tratam-se de projetos inovadores, quer por usarem técnicas e algoritmos que podem não estar consolidadas ou na Organização ou em pesquisas acadêmicas, quer por envolverem a construção de modelos que simulam processos cognitivos, inteligência natural, por máquinas.

São projetos quase sempre únicos. As particularidades de cada contexto, dos dados envolvidos, dos requisitos de qualidade, impedem ou dificultam o reaproveitamento de código para outros projetos.

São projetos complexos pois as técnicas empregadas em geral têm concepção de difícil entendimento e envolvem conhecimento interdisciplinar de áreas como ciência da computação, matemática e estatística, além do entendimento do negócio para o qual a solução se destina.

São processos exploratórios, pois a atividade de mineração de dados pode ser definida como o processo de *explorar* um conjunto de dados, com técnicas diversas, extraindo ou ajudando a evidenciar padrões e auxiliando na descoberta de conhecimento.

E, no caso de organizações com baixo grau de maturidade em DM, são projetos que ficam sob a responsabilidade de pequenas equipes ou mesmo de um único indivíduo. Neste caso, a partilha do conhecimento e de práticas adotadas nos projetos fica ainda mais difícil.

Infelizmente, esses trabalhos complexos, de caráter exploratório, inovadores, únicos, e, em geral, individuais não deixam rastro do que fazem[1]. Ao final temos a solução implementada. Pode-se ter até uma documentação do produto criado, mas não do processo seguido, das escolhas feitas e das técnicas usadas nas diversas atividades do projeto. Resta um temor no ar para o caso de ser necessário reconstruir o modelo sem a presença do seu criador. E o responsável se torna pai do produto, pois só ele o conhece e pode manter.

Uma questão intrigante na gestão do conhecimento em geral é como coletar, colher ou tornar explícita a experiência de projetos para que possam ser utilizáveis para outros

[1] Essa situação caótica tende a não ocorrer em organizações maduras em DM, que usam análise de dados intensamente e criam equipes ou mesmo unidades dedicadas a projetos de DM.

(DINGSØYR et al, 2001). Pois a memória de uma organização não pode se basear apenas na memória de seus indivíduos (STATA, 1980).

Diante do exposto surge a desafiadora questão de pesquisa: como sistematizar a documentação das tarefas de um projeto de mineração de dados de forma a potencializar sua auditabilidade e a partilha dos aprendizados?

Buscando responder à questão colocada, o objetivo deste trabalho é propor um conjunto de boas práticas de registro semi-automatizado das atividades de um projeto de DM de forma a deixar um rastro das escolhas feitas, dos processamentos realizados e dos resultados obtidos, com foco não no produto gerado, mas no processo por trás de sua construção. Boas práticas que possam ser mescladas à metodologia corporativa de DM em uso na organização.

A produção de rastros em projetos de DM pode acelerar a curva de aprendizagem e a formação de uma cultura organizacional em torno do uso da análise de dados.

São três os objetivos específicos: contextualização do referencial teórico sobre metodologias e documentação em projetos de mineração de dados, bem como do potencial impacto das experiências adquiridas nos projetos para uma organização; proposição do Rastro-DM com a descrição de suas atividades e a ilustração de sua aplicação em um projeto de classificação textual de documentos em PDF associados a danos ao Erário Público Federal Brasileiro denominado Cladop, para avaliação de sua viabilidade e dos benefícios de seu uso. Os objetivos específicos serão tratados nas três seções que se seguem.

2 REFERENCIAL TEÓRICO

2.1 METODOLOGIAS DE MINERAÇÃO DE DADOS

Segundo BERMAN *et al* (2018), a Ciência de Dados se concentra nos processos de extração de conhecimento ou *insights* a partir de dados estruturados ou não. Esse processo de descoberta de conhecimento em dados (KDD - *Knowledge-Discovery in Databases*), segundo BECKER e GHEDINI (2005), é complexo e popularmente chamado de Mineração de Dados (DM - Data Mining). No escopo deste trabalho chamaremos indistintamente KDD de DM.

WIRTH e HIPP (2000) afirmam que DM é um processo complexo. Associam o sucesso de um projeto[2] de mineração de dados a uma adequada combinação de boas ferramentas, analistas qualificados, uso de uma metodologia sólida e um eficaz gerenciamento

[2] No contexto deste trabalho, usaremos o termo projeto para indicar projeto de mineração de dados.

de projetos. Quanto à metodologia, os mesmos autores afirmam que DM precisa de uma abordagem padrão que ajude a transformar problemas de negócios em tarefas de mineração de dados, que sugira transformações de dados apropriadas e técnicas a empregar, e forneça meios para avaliar a eficácia dos resultados e documentar a experiência. E destacam que o uso de uma metodologia no planejamento e na apresentação de relatórios inspira confiança nos usuários e nos patrocinadores.

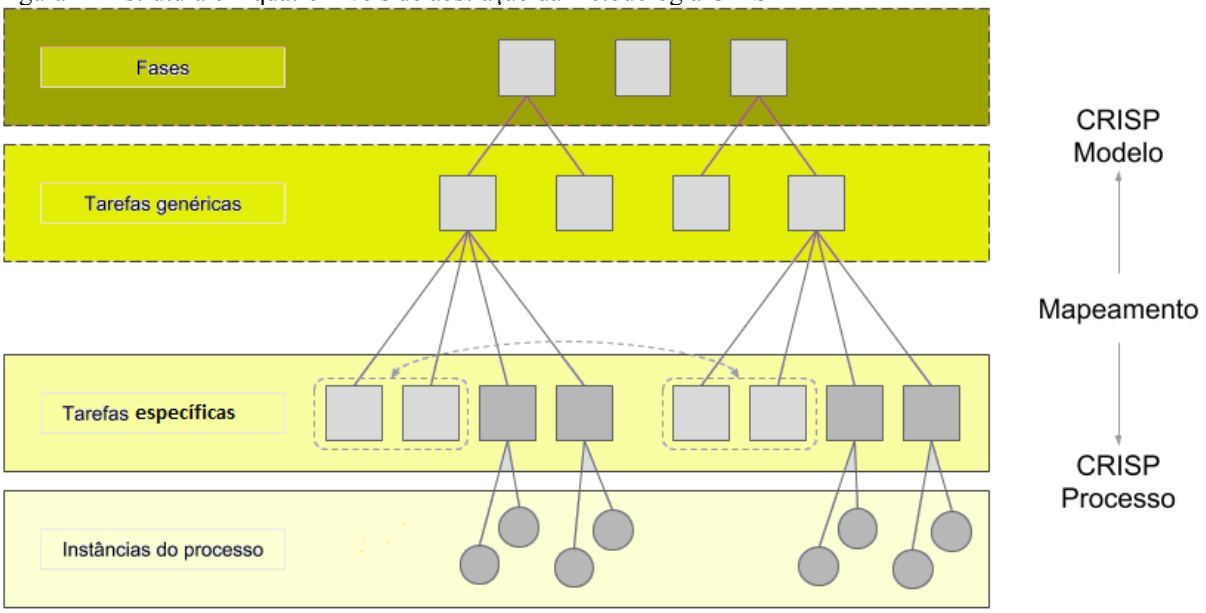
KURGAN e MUSILEK (2006) reforçam a melhor aceitação de projetos que usam metodologias e acrescentam uma melhor compreensão e entendimento do processo. E vão além, ao destacar que elas promovem economia de tempo e de custos com o estabelecimento de um roteiro a seguir para o planejamento e a execução dos projetos. Mas alertam que um grande número de projetos segue metodologias próprias.

MINGERS e BROCKLESBY (1997) relatam que há várias interpretações possíveis para os termos paradigma, metodologia, método e técnica e, após apresentarem a interpretação deles, correlacionam os conceitos: uma metodologia especifica que tipo de atividades devem ser realizadas, as técnicas são formas particulares de realizar essas atividades e os paradigmas trazem as filosofias que motivam as atividades nas metodologias. Evitam o uso do termo *método* por ter uma ambiguidade: às vezes usado como sinônimo de técnica e outras de metodologia.

Um exemplo de metodologia usada em DM é *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Ela é considerada padrão de fato e um dos fatores de seu sucesso é ser neutra em termos de indústria, ferramentas e aplicações (MARISCAL et al, 2010). Os mesmos autores afirmam que uma metodologia não só deve especificar as tarefas, suas entradas e saídas, mas também a maneira como devem ser executadas.

CHAPMAN *et al* (2000) apresentaram a metodologia CRISP-DM como um modelo de processo hierárquico, consistindo em conjuntos de tarefas descritas em quatro níveis de abstração (do geral ao específico): fase, tarefa genérica, tarefa especializada e instância do processo (ver Figura 1). Os autores detalham ao final de sua obra os principais grupos de problemas tratados pela DM (com algumas propostas de técnicas apropriadas à época): descrição de dados e sumarização, segmentação (clusterização), descrição de conceitos, classificação, predição (regressão) e análise de dependência.

Figura 1 - Estrutura em quatro níveis de abstração da metodologia CRISP-DM

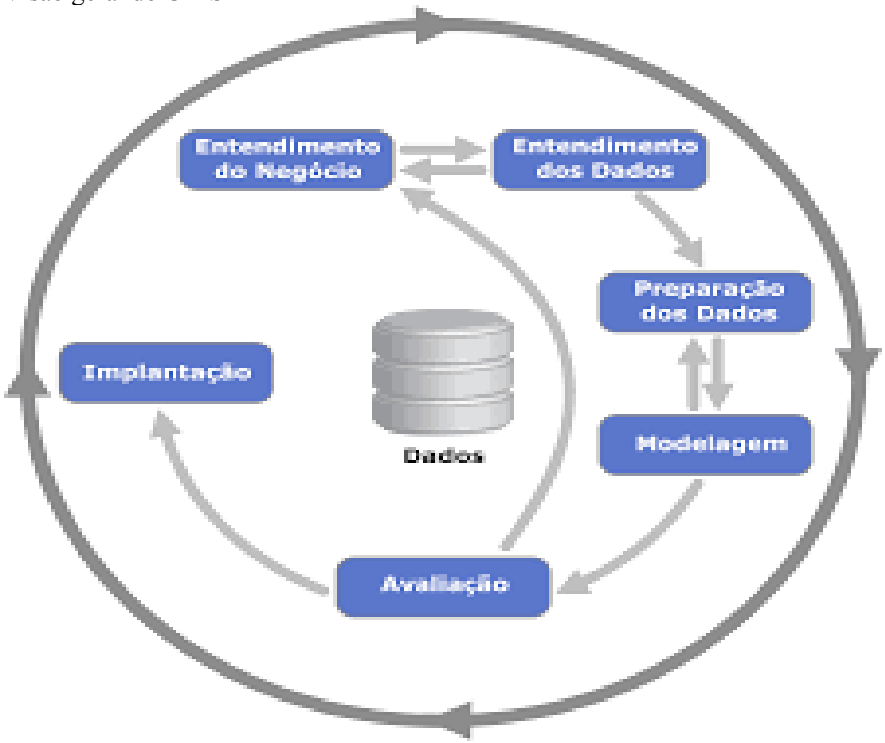


Fonte: CHAPMAN et al (2000)

A Figura 2 mostra a iteração entre as fases do CRISP-DM[3]. CHAPMAN *et al* (2000) afirmam que a sequência das fases não é rígida. Que, na prática, muitas das tarefas podem ser executadas em uma ordem diferente, e, muitas vezes, faz-se necessário voltar repetidamente para tarefas anteriores e repetir certas ações. Afirmam que representar todas as rotas possíveis por meio do processo de mineração de dados exigiria um modelo de processo excessivamente complexo, por isso não o fazem. Afirmam que nunca uma fase é completamente concluída antes que a fase seguinte comece.

[3] Sugere-se, como referência visual, o gráfico que engloba as atividades e as saídas de cada fase do CRIP-DM e pode ser encontrado em: LEAPER, Nicole. *A visual guide to CRISP-DM methodology*. Site: **https://exde.files.wordpress.com/2009/03/crisp_visualguide.pdf** [Acesso em 27/7/2019], 2009

Figura 2 - Visão geral do CRISP-DM



Fonte: CHAPMAN *et al* (2000)

Segundo MARBÁN *et al* (2007), o CRISP-DM não cobre muitas tarefas relacionadas à gestão de projeto, organização e qualidade. Pelo menos não da maneira exigida pela crescente complexidade dos recentes projetos de mineração de dados que envolvem não apenas grandes volumes de dados, mas também o gerenciamento e a organização de grandes times interdisciplinares.

BECKER e GHEDINI (2005) acentuam que a estrutura de um processo de DM é altamente dependente da metodologia adotada, das habilidades, da experiência e do estilo da pessoa responsável pelo processo, bem como dos recursos disponíveis na corporação. E confirmam a alta iteratividade e interatividade dos processos. Ressaltam que embora a estrutura conceitual do processo sugira uma ordem entre as fases, na prática, os analistas passam de qualquer fase para quase qualquer outra fase a qualquer momento, até porque muitos problemas relacionados a fases anteriores (por exemplo, preparação de dados) só podem ser detectados muito mais tarde, quando os padrões e os modelos são avaliados.

MARBÁN *et al* (2007) agrupam as tarefas de um processo de DM, em relação à construção de um modelo, em três estágios: pré-desenvolvimento, desenvolvimento e pós-desenvolvimento. Afirmam que todas as metodologias usadas para DM se concentram no estágio de desenvolvimento, que equivale a coleta e a análise dos dados disponíveis para o projeto, a criação de novos dados a partir dos disponíveis, a adaptação para algoritmos de DM e a criação de modelos. Segundo CHOLLET (2017), o desenvolvimento de um projeto está

centrado em experimentações de um modelo: você começa com uma ideia e a expressa como um experimento, tentando validar ou invalidar sua ideia. Depois, executa-se essa experiência e processam-se as informações geradas. Isso inspira sua próxima ideia. Quanto mais iterações desse círculo repetitivo você conseguir executar, mais refinadas e poderosas suas ideias se tornarão. A experimentação, ou seja, a aplicação de algoritmos matemáticos aos dados para a extração de padrões, é chamada de treinamento (*training*) por NGUYEN (2018). É importante obter o máximo possível de informações dos treinamentos, incluindo o desempenho dos modelos (CHOLLET, 2017). Para simplificar, no contexto deste trabalho, será usado o termo *treinamento* e *modelo* representando a experimentação e o produto resultante dos padrões detectados nos dados.

GREFF *et al* (2017) confirmam o número significativo de experimentos computacionais com muitas configurações diferentes de hiperparâmetros e alertam sobre o desafio prático da documentação. Segundo os autores, devido à pressão de prazo e a inerente natureza imprevisível de um projeto de DM, geralmente há pouco incentivo para a construção de infraestruturas robustas e, como resultado, o código muitas vezes evolui rapidamente, o que compromete, entre outras coisas, a documentação do projeto.

Segundo BECKER e GHEDINI (2005), os projetos são desenvolvidos em geral de maneira não estruturada, *ad hoc*: após a análise inicial dos dados, decide-se experimentar uma determinada técnica, cujos resultados podem sugerir a reestruturação dos dados e a execução de novos tipos de análises; e assim por diante. E, à medida que o tempo passa, não se pode lembrar quais treinamentos foram realizados, os conjuntos de dados utilizados, os hiperparâmetros usados e, mais importante, os resultados que foram derivados dos conjuntos de dados. E essa situação, segundo as autoras, leva à reexecução de treinamentos. Alertam que a situação é ainda pior se forem considerados projetos de longo prazo envolvendo várias pessoas. Constatam que, apesar da diversidade de conhecimento das pessoas, das técnicas e das ferramentas, a maioria dos projetos de DM enfrenta, na prática, as mesmas dificuldades: as experimentações com parâmetros, as transformações de dados, o desperdício de se refazer um trabalho e o gerenciamento de recursos e de resultados. Afirmam as autoras que a documentação do histórico das tarefas face a iteratividade e a interatividade do processo é um problema aberto no gerenciamento de projetos de DM.

GHEDINI e BECKER (2001) constatam que, apesar do seu valor, a documentação é mais frequentemente percebida como uma sobrecarga, cujos benefícios são reconhecidos tardiamente. Embora reconheçam que o CRISP-DM esteja fortemente acoplado à atividade de documentação (de decisões e de suas razões, de suposições, de descobertas e de conhecimentos

adquiridos na execução do projeto), criticam o fato de a metodologia não abordar explicitamente a iteratividade inerente ao processo, nem como os redirecionamentos e redefinições afetam e atualizam a documentação produzida. O CRISP-DM também não fornece uma estrutura para relatórios e saídas, meramente descrevendo seu propósito e o conteúdo geral (BECKER e GHEDINI, 2005).

2.2 CONHECIMENTO EM PROJETO DE DM: UM POTENCIAL SALTO ORGANIZACIONAL

Enquanto os dados podem ser adquiridos e conectados cada vez mais facilmente, a conexão entre projetos a partir da colaboração de competências técnicas não existe (HUBER et al, 2018, apud HUBER et al, 2019).

WIRTH e HIPP (2000) salientam que o sucesso ou o fracasso de um projeto de mineração de dados é altamente dependente da pessoa ou da equipe e que práticas de sucesso não são necessariamente repetidas em toda a empresa.

Para direcionar o conhecimento individual para os propósitos organizacionais, uma organização deve desenvolver e nutrir um ambiente de compartilhamento de conhecimento, transformação e integração entre seus membros (NONAKA E TAKEUCHI, 1995, apud BHATT, 2001).

CHAPMAN *et al* (2000) ressaltam que projetos de mineração de dados podem ser beneficiados com as experiências de projetos anteriores. As lições aprendidas durante o processo e a partir da solução implantada podem desencadear novas questões de negócios, muitas vezes mais focadas.

BECKER e GHEDINI (2005) exemplificam benefícios de se partilhar conhecimento entre projetos citando que experiências de projetos anteriores podem ser usadas para estabelecer planos de projetos de DM mais razoáveis, com estimativas mais precisas de cronogramas, orçamento, etc. A experiência torna mais fácil defender recursos mais realistas, pois há uma compreensão mais profunda de como o esforço é realmente gasto. Além disso, pode-se usar experiências anteriores para lidar com certas classes de problemas ou de técnicas. Defendem que a documentação da execução de projetos deve ser tratada como um recurso corporativo, que pode ser compartilhado pela equipe, ser usado como referência e pode estar sujeito a políticas e padrões corporativos.

Segundo BHATT (2001), à medida que os indivíduos nas organizações interagem com outros (indivíduos, tecnologias e técnicas), eles tendem a entender e partilhar suas visões

diferentes sobre as mesmas situações, construindo suas comunidades e compartilhando técnicas eficientes de trabalho. Segundo os autores, esse processo de interação é útil para desenvolver uma visão holística das realidades, facilitando a integração de um corpo diversificado de conhecimentos nas organizações.

É fato que as empresas precisam educar um número maior de pessoas sobre os processos e as melhores práticas associadas de mineração de dados e de análise preditiva (MARISCAL et al, 2010). Sabe-se também que através de processos e práticas, o conhecimento pode ser incorporado em indivíduos ou em organizações (NGUYEN, 2018). Urge a necessidade de se alavancar o conhecimento organizacional com a partilha dos conhecimentos entre os projetos de DM.

BHATT (2001) afirma que o conhecimento organizacional é formado por padrões únicos de interações entre tecnologias, técnicas e pessoas, que não podem ser copiados facilmente por outras organizações, porque essas interações são únicas da organização, moldadas pela sua história e pela sua cultura. O mesmo autor credita a sustentação de vantagens competitivas da empresa no longo prazo ao incentivo ao crescimento desse conhecimento com a criação de um ambiente estimulante e prático (*aprender-fazendo*).

NGUYEN (2018) enfatiza que o conhecimento é uma *commodity* cara para as organizações, que vem de muitas fontes, como, por exemplo, documentos, processos, pessoas, comunicação, cultura e aprendizagem. E que a transferência de conhecimento estimula inovações reforçando a compreensão dos indivíduos e aumentando a quantidade de conhecimento para cada pessoa.

BHATT (2001) traz definições precisas de conhecimento, informação e dados. Em geral, os dados são considerados fatos brutos, a informação é considerada como um conjunto organizado de dados e o conhecimento é percebido como uma informação significativa. O conhecimento é uma combinação organizada de dados, assimilados por um conjunto de regras, procedimentos e operações aprendidas por meio da experiência e da prática. Sem significado, conhecimento é informação ou dados. A distinção entre informação e conhecimento é subjetiva e depende das perspectivas dos usuários na capacidade de atribuir significado à informação. Sendo o ciclo entre dados, informação e conhecimento recursivo, uma organização deve ser rápida para transformar dados em informação e informação em conhecimento, defende o mesmo autor.

A criação de conhecimento pode se dar pelo uso de tecnologia, pela condução de experimentos e pela elaboração de síntese de informações (APPEL-NASA-GOV, 2015). DINGSØYR *et al* (2001) tratam de forma indistinta conhecimento e experiência. Embora

reconheçam que experiência em um sentido estrito é algo que reside nos seres humanos e que não pode ser transferido para os outros (que teriam que experimentar por si mesmo para ter a experiência), em uma definição menos estrita, afirmam que experiência é informação que é *operacional*, isto é, utilizável em alguma situação. Entendem que uma descrição de um evento acontecido em um projeto é um item de experiência.

Como comentado na introdução, uma questão interessante na gestão do conhecimento em geral é como coletar, colher ou tornar explícita a experiência de projetos para que possam ser utilizáveis para outros (DINGSØYR et al, 2001).

Como possível resposta, encontramos em (NGUYEN, 2018) o uso de processos de gestão do conhecimento (KMP - *Knowledge Management Process*), que objetivam fazer circular o conhecimento em toda a organização para garantir que o conhecimento certo chegue à pessoa certa para entender e ter conhecimento suficiente para tomar decisões e executar bem as tarefas. Destaca o autor que KMP pode ser usado em qualquer nível, desde a organização como um todo até uma equipe, passando por departamentos e complementa que seus estágios (identificação, criação, armazenamento, transferência e utilização de conhecimento) estão interligados e são iterativos, uma vez que o conhecimento é continuamente formado e mudado.

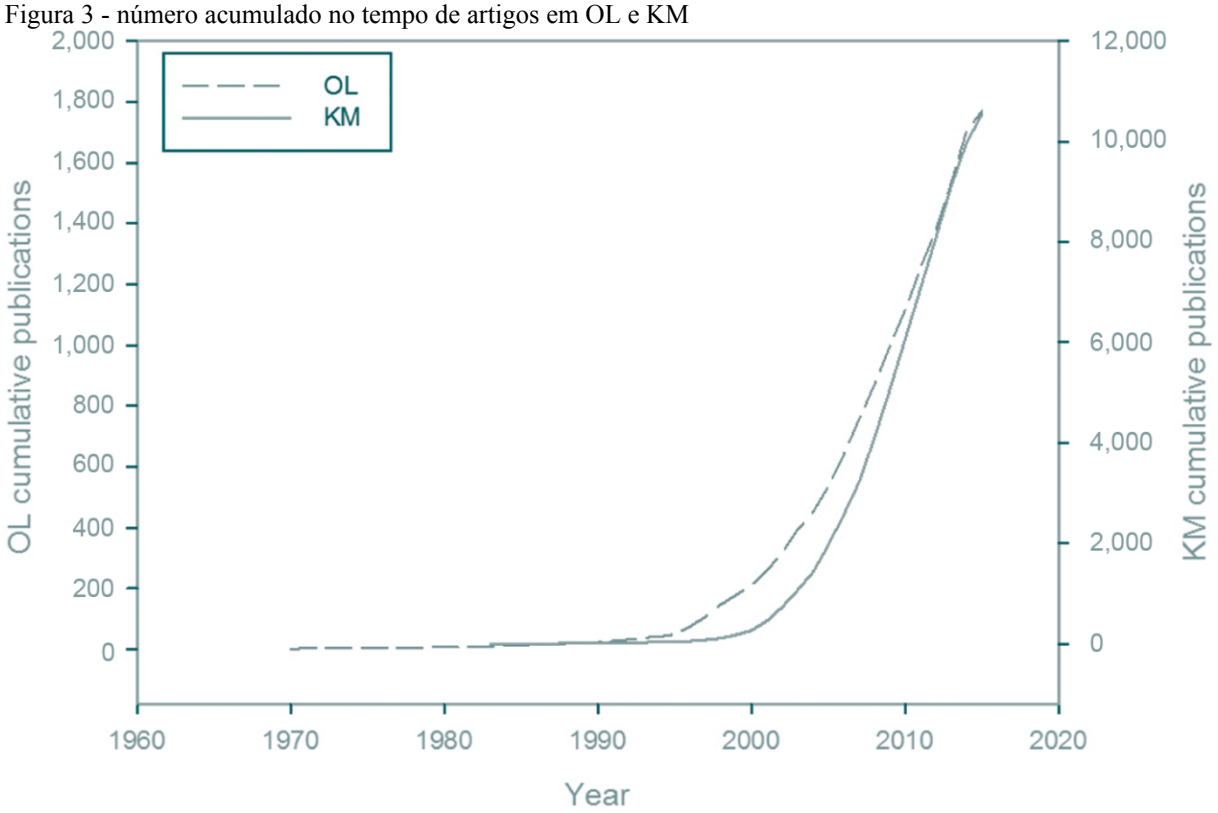
O mesmo autor encontra na combinação entre KMP e DM um grande potencial na exploração e no gerenciamento do conhecimento valioso de big data (DM suporta o KMP na geração de conhecimento inestimável e KMP suporta DM na coleta e armazenamento de conhecimento como entrada de DM), mas ressalta que há um grande vazio de pesquisas nessa área.

STATA (1980) traz o conceito de aprendizado organizacional (OL - *Organizational Learning*), que ocorre por meio de *insights* compartilhados, conhecimento e modelos mentais e que se baseia no conhecimento e na experiência do passado, ou seja, na memória. O autor detalha que a memória organizacional depende de mecanismos institucionais (por exemplo, políticas e estratégias) usados para reter o conhecimento, não podendo depender exclusivamente da memória dos indivíduos, pois há sempre o risco de se perder lições e experiências duramente conquistadas à medida que as pessoas migram de um emprego para outro. Além de outros motivos de saída de pessoas (aposentadoria, afastamentos, transferências, falecimento, etc.), podemos acrescentar também o risco do esquecimento.

Entre outros benefícios, o aprendizado organizacional fomenta a inovação, uma vez que esta é bloqueada a menos que todos os principais tomadores de decisão aprendam juntos, compartilhem crenças e objetivos e estejam comprometidos em tomar as ações necessárias para a mudança (STATA, 1980).

CASTANEDA *et al* (2018) conduziram uma pesquisa sistemática e comparativa sobre os temas OL e gestão de conhecimento (KM - *Knowledge Management*) que surgiram respectivamente em 1963 e 1993. A Figura 3 apresenta o número acumulado no tempo de artigos em cada temática. Percebe-se um crescimento nos últimos anos. Eles concluem que KM está, de certa forma, absorvendo OL e propõem estudos futuros que investiguem o papel do compartilhamento de conhecimento como um conector entre KM e OL. Sobre esse conector, eles afirmam que indivíduos e organizações aprendem compartilhando conhecimento e que, de acordo com a literatura de KM, o compartilhamento de conhecimento é o processo central responsável pela criação e aplicação do conhecimento.

Figura 3 - número acumulado no tempo de artigos em OL e KM



Fonte: CASTANEDA et al (2018)

O próprio CRISP-DM reforça a necessidade do registro de experiências. Como pode ser visto em sua documentação (CHAPMAN et al. 2000), a atividade *Revisar projeto* tem como saída um documento intitulado *Documentação de Experiência* e é descrita como: *Resuma a importante experiência adquirida durante o projeto. Por exemplo, armadilhas, abordagens enganosas ou dicas para selecionar as técnicas de mineração de dados mais adequadas em situações semelhantes poderiam fazer parte dessa documentação*.

Segundo BECKER e GHEDINI (2005), com a documentação do processo de DM, à medida que o conhecimento é tornado explícito e gerenciado, ele aumenta o intelecto da organização, tornando-se uma base para a comunicação e para a aprendizagem, apoiando a disseminação de conhecimento e a experiência dentro da organização em vários níveis. As autoras reforçam que a ideia de capturar e armazenar todo o conhecimento informal relevante gerado e usado durante um processo de DM, de modo que esteja disponível para recuperação posterior, constitui uma abordagem interessante para lidar com a dificuldade citada de refletir na documentação a iteratividade e a interatividade do processo.

NGUYEN (2018) afirma que armazenar conhecimento é o caminho para se criar uma propriedade inestimável para as organizações. Um bem que se acumula ao longo do tempo e que não pode ser comprado por dinheiro algum.

Diante do exposto, podemos concluir que a documentação da memória de um projeto é um pequeno passo, uma vez que localizado em um contexto menor e a um custo menor, que pode significar um salto em direção à memória organizacional e aos ganhos derivados do aprendizado partilhado e do conhecimento gerenciado.

2.3 DOCUMENTAÇÃO: CAMINHO PARA A GERAÇÃO DE CONHECIMENTO PARTILHÁVEL

GHEDINI e BECKER (2000) ressaltam que a documentação de experimentos e de todas as partes relevantes de um projeto não só evita a perda de conhecimento confinado nas mentes das pessoas como também permite o seu compartilhamento, tornando-se uma rica fonte de conhecimento para referência futura e reuso corporativo. Destacam também que essa atividade leva a um melhor gerenciamento de esforços, de recursos e de resultados de um projeto de DM.

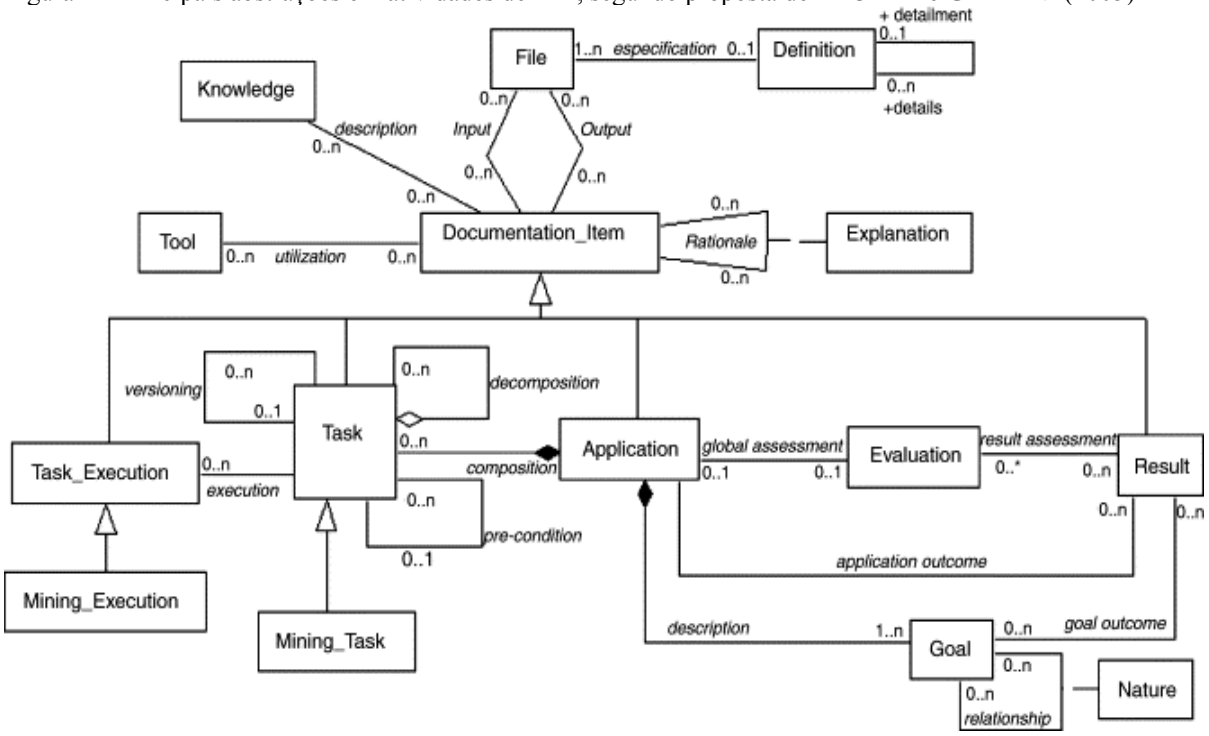
Segundo PRAKASH *et al* (2012), informações de treinamentos, código de implementação e seu histórico de mudança contêm uma riqueza de informações sobre o estado, o progresso e a evolução de um projeto de software. Afirmam também que a mineração de dados está se tornando uma ferramenta cada vez mais importante para transformar esses dados em informações. Por analogia, espera-se que dados de projetos de DM sejam convertidos também em informações por atividades de mineração de dados.

WIRTH e HIPP (2000) enfatizam que, talvez, o maior benefício de terem aplicado uma metodologia foi a documentação gerada. Admitem terem pulado inicialmente algumas tarefas de documentação e planejamento por serem demoradas e por considerarem desnecessárias para especialistas como eles. Mas apresentam o preço que pagaram por essa

ação e se arrependem constatando que todo esforço vale a pena. Relatam alguns benefícios observados a partir da documentação produzida: evita o desperdício de esforço (por exemplo, em caminhos não frutíferos ou com trabalho repetitivo); promove um gerenciamento eficaz e uma melhor comunicação da equipe; permite a identificação de pontos críticos no processo; promove um melhor planejamento de projetos futuros, com base em uma melhor percepção de como o esforço foi gasto e dos recursos necessários; promove o uso de experiências documentadas em outros contextos.

BECKER e GHEDINI (2005) também identificam o papel da documentação na aprendizagem e na reutilização e afirmam que um benefício imediato da documentação é a efetividade no gerenciamento, no planejamento e na comunicação. Constatam que a documentação é completamente dependente da equipe do projeto, uma vez que está em sua responsabilidade a veracidade dos registros e o nível de detalhe que impacta diretamente sua utilidade. Quanto à resistência, afirmam que, à medida que os benefícios da atividade são percebidos, há um estímulo a documentar o processo com mais detalhes e de forma concomitante com a execução das atividades e que os melhores resultados são alcançados a longo prazo, quando a equipe do projeto descobre qual estratégia melhor se adequa ao seu estilo de trabalho, bem como as melhores maneiras de obter vantagens dos recursos e das técnicas e da flexibilidade do modelo. E propõem uma infraestrutura de documentação composta de um modelo de documentação e um ambiente de suporte que permite a captura, armazenamento e recuperação de informações e artefatos relacionados ao processo de DM. O modelo de documentação proposto fornece um conjunto de abstrações específicas relacionadas às atividades de DM, suas entradas e suas saídas, e está representado em forma de diagrama de classe UML na Figura 4. Uma lamentação das autoras quanto ao protótipo usado foi a necessidade de se sair da plataforma em uso de exploração do modelo para se efetuarem os registros.

Figura 4 - Principais abstrações em atividades de DM, segundo proposta de BECKER e GHEDINI (2005)



Fonte: BECKER e GHEDINI (2005)

DINGSØYR *et al* (2001) defendem que seja elaborado ao final de um projeto um *Relatório de Experiência* para coletar o que deu certo e o que deu errado no processo adotado[4].

Contudo, WIRTH e HIPP (2000) alertam sobre a dificuldade de se documentar ao final, de se tentar reconstruir o que foi feito e suas motivações. Enfatizam que os processos de DM são vivos e, como tal, a documentação deve ser flexível e viva, e não deve ser atualizada após o final do projeto (*post mortem*). Defendem que a definição de uma estratégia de documentação deve ser um ponto de partida, mas a flexibilidade para a evolução e a mudança deve ser uma premissa. Enfatizam que encontrar o nível certo de detalhes para se planejar e se documentar um processo de DM é difícil e faz parte de um longo processo de aprendizado, e pode ser influenciado por diversos fatores como complexidade do projeto, duração e tamanho da equipe.

GHEDINI e BECKER (2000) defendem que a equipe do projeto defina uma estratégia de documentação (pelo menos inicial) antes do início do processo, o que seria sua própria metodologia de documentação. Afirmam que o CRISP-DM distingue dois tipos de documentação: um direcionado às pessoas envolvidas na atividade como meio de orientar o processo e gerenciar os esforços e outro destinado às pessoas não envolvidas nas atividades,

[4] O que já é previsto na atividade *Revisar Projeto* do CRISP-DM, que produz um documento intitulado *Documentação de Experiência* (CHAPMAN et al, 2000).

muito mais resumida, pois se concentra apenas em comunicar os resultados significativos (principalmente relacionados a atividades bem-sucedidas).

GREFF *et al* (2017) citam que uma dificuldade para o compartilhamento e a colaboração e a reprodutibilidade dos experimentos é o uso pelas equipes de configurações particulares de área de trabalho no uso das diferentes ferramentas necessárias para tratar os aspectos distintos de um processo de DM, citando, entre outras, bancos de dados, sistemas de controle de versão, ferramentas automatizadas de otimização de hiperparâmetros, scripts e planilhas.

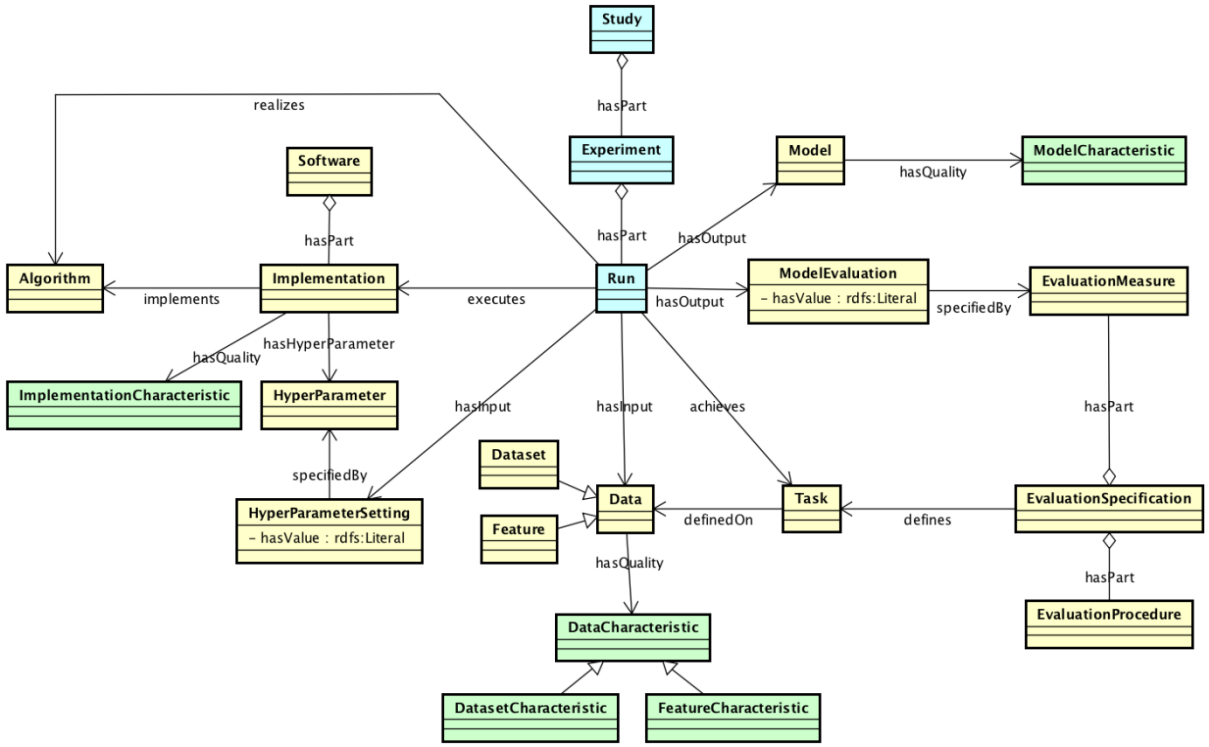
PRAKASH *et al* (2012) relatam que numerosos trabalhos estão sendo feitos no desenvolvimento de plataformas integradas para *Machine Learning* (ML) e para Engenharia de Software baseadas em componentes reutilizáveis, citando, entre as de código aberto mais conhecidas, WEKA e Rapid Miner.

GREFF *et al* (2017) apresentam *Sacred*, um framework *Python* de código aberto que visa fornecer infraestrutura básica para a execução de experimentos computacionais de forma independente dos métodos e bibliotecas utilizados. Concentram-se em resolver problemas como o gerenciamento de configurações, a documentação e a reprodutibilidade dos resultados. Para cada treinamento, informações relevantes, como parâmetros, dependências de pacotes, informações do *host*, código-fonte e resultados, são capturadas automaticamente e armazenadas em um repositório centralizado, de onde pode-se consultar também detalhes dos hiperparâmetros usados e dos resultados obtidos.

PUBLIO *et al* (2018) acreditam que a visão de modelos canônicos e padronizados pode levar a uma melhor compreensão dos dados e dos algoritmos de ML empregados em DM e pode promover a interoperabilidade dos experimentos, independentemente da plataforma ou da solução de fluxo de trabalho adotada.

W3C Machine Learning Schema Community Group (2017) publicou uma ontologia *machine learning schema* (ML Schema) que fornece um conjunto de classes, propriedades e restrições para representar e intercambiar informações sobre algoritmos de aprendizado de máquina, conjuntos de dados e experimentos. Segundo o grupo, a ontologia pode ser facilmente estendida e mapeada para outras ontologias mais específicas de domínio desenvolvidas na área de aprendizado de máquina e mineração de dados. A Figura 5 traz a visualização das principais abstrações do ML Schema.

Figura 5 - Conceitos principais do ML Schema



Fonte: W3C Machine Learning Schema Community Group (2017)

KURGAN e MUSILEK (2006) afirmam a possibilidade de integração e de interoperabilidade dos modelos de DM com o uso de padrões industriais como PMML (*Predictive Model Markup Language*) que representa um modelo em um esquema XML (*Extensible Markup Language*). Segundo os autores, pode-se usar ferramentas diferentes para a geração, visualização e análise de um mesmo modelo.

3 RASTRO-DM: CONJUNTO DE BOAS PRÁTICAS

3.1 VISÃO GERAL

O Rastro-DM é um conjunto de boas práticas propostas neste trabalho que objetiva a documentação das definições de ação, dos treinamentos de modelo e dos aprendizados concebidos em um projeto de mineração de dados. Engloba três atividades[5] que correspondem aos conceitos documentados:

[5] Considerando o significado usual de rastro, *Percurso ou comportamento de alguém, que pode ser seguido ou imitado*, podemos pensar em uma analogia para uma melhor compreensão dos conceitos. Se o objetivo do projeto fosse procurar em uma floresta por uma pedra que mais se assemelhasse a um conjunto de dados (características como volume, grau de transparência, peso, altura, etc.), as placas indicativas dos caminhos seguidos pelo aventureiro corresponderiam às Definições de Ação: *vire à direita na figueira*, *atravesse o riacho*

- Definição de ação;
- Registro de treinamento;
- Síntese de aprendizado.

As atividades do Rastro-DM são complementares às tarefas previstas na metodologia em uso na organização, aqui chamada de *metodologia base*, que traz todo o arcabouço metodológico e paradigmático usado em um processo de DM, qualquer que seja ela. O fato de ser adaptável a várias metodologias de DM permite a quebra da resistência das pessoas e o ceticismo em relação à documentação do processo (GHEDINI e BECKER, 2001). Por padronização e clareza, será usado nas explicações o CRISP-DM como *metodologia base,* e os passos da *metodologia base* serão referenciados como *tarefas CRISP-DM* e os do Rastro-DM como *atividades Rastro-DM*.

MINGERS e BROCKLESBY (1997) identificam várias maneiras de se combinar metodologias. Afirmam que o estabelecimento de boas práticas é uma forma de criação de uma nova metodologia, logo, não há incongruência em se referir ao Rastro-DM como uma metodologia.

As atividades Rastro-DM ocorrem várias vezes durante um projeto e podem estar associadas a uma ou mais tarefas CRISP-DM. Por exemplo, a síntese de um aprendizado como em[6]: *Ao contrário dos algoritmos shallow, redes alcançam uma melhor performance se as variáveis categóricas forem transformadas em colunas dummies* referencia pelos assuntos tratados, formatação de dados e algoritmos, mais de uma tarefa CRISP-DM: *Formatar dados* e *Selecionar técnica*.

No Rastro-DM, as atividades não devem ser executadas ao final do projeto, mas de forma concomitante ao avanço dos trabalhos para que a documentação gerada seja útil também ao projeto em andamento. Afinal, a documentação tempestiva é mais rica em informações e ajuda a conduzir o projeto de forma mais eficaz (GHEDINI e BECKER, 2000) e uma documentação *post mortem*, atualizada apenas após o final do projeto, não se adequa a projetos de DM, que são considerados *vivos* dada sua complexidade e iteratividade (WIRTH e HIPP, 2000).

---

na pequena ilha, etc. As verificações sobre o quanto as pedras encontradas pelo caminho se encaixam aos dados, seriam os Treinamentos: com o registro de fotos das pedras, sua localização, e os dados apurados, altura, grau de transparência, etc. O conhecimento construído durante o projeto seriam os Aprendizados, como, por exemplo: *atenção ao risco de cobras próximo a riachos* ou *é importante lavar a pedra com detergente neutro antes de tomar suas medidas*. .

[6] Aprendizado hipotético usado para ilustração.

Como visto, capturar e armazenar todo o conhecimento informal relevante gerado e usado durante um processo de DM, para recuperação posterior, constitui uma abordagem interessante para lidar com a dificuldade de refletir na documentação a iteratividade e a interatividade do processo (BECKER e GHEDINI, 2005).

Rastro-DM é flexível ao não definir uma relação mínima de atributos de cada conceito. Afinal cada projeto e organização tem sua complexidade particular e seu grau de amadurecimento em DM. Em última análise, cabe à equipe não só a veracidade e a efetividade da documentação como também a definição do quê documentar e quando documentar, bem como o nível de detalhe da documentação[7].

A seguir, serão descritas as atividades do Rastro-DM.

3.2 DEFINIÇÃO DE AÇÃO

Atividade responsável por documentar a definição dos passos de um projeto, executados ou não. Conceitualmente, uma definição de ação corresponde à descrição de uma ou mais tarefas específicas do CRISP-DM (instanciadas em um processo, conforme Figura 1).

O objetivo da atividade não é o registro dos detalhes da execução de uma ação, mas as informações sobre sua definição, como a declaração de seu objetivo e as técnicas a serem usadas ou experimentadas quando da execução da ação[8].

A atividade de definição de ação pode ocorrer a qualquer momento de um projeto de DM.

Como visto, segundo BECKER e GHEDINI (2005), a maioria dos projetos de DM enfrenta, na prática, dificuldades de gerenciar recursos e resultados, atestando que a documentação do histórico dos passos, face a iteratividade e a interatividade do processo, é um problema aberto no gerenciamento de projetos de DM.

É importante a definição de cada ação que se inicia, pois, de certa forma, justifica os treinamentos que se seguirão. Além disso, saber o que já foi feito impedirá a perda de esforço com execuções repetidas.

---

[7] Espera-se que, com o tempo e com o amadurecimento do processo de DM nas organizações, seja elaborado um padrão corporativo mínimo de documentação por conceito e por tarefa de DM. Mas, esse padrão não pode afastar a criatividade da equipe e não se tornar apenas uma sobrecarga de trabalho.

[8] Detalhes da execução da ação definida podem ser registrados, se desejado, como aprendizados do projeto relacionados à ação definida.

O quê e quando registrar, bem como o nível de abstração e de detalhe, se mais próximo de uma tarefa CRISP-DM, como, *formatação de dados*, ou mais detalhada, como, *formatação do nome do arquivo para possível retirada de stopwords e de pontuações*, cabe a cada equipe[9].

Diante das dificuldades de gerenciamento em processos de DM identificadas no referencial teórico, pode-se avaliar a possibilidade do registro de recursos usados nas ações definidas (pessoas, tempo, etc.). Essas informações se forem sobre ações futuras, podem apoiar estimativas de prazo e o estabelecimento de cronogramas do projeto. De qualquer forma, o histórico desses dados pode servir de base para alocação de recursos em projetos futuros.

Embora a simplicidade de um campo textual descritivo seja aceitável, uma vez que o conteúdo é o mais importante, quanto mais estruturado for o registro, melhor traduzirá a compreensão do processo e do projeto em andamento e maior será a contribuição para o resultado final. Essa visão é defendida também por PUBLIO *et al* (2018) que enfatizam que a visão de modelos canônicos e padronizados pode levar a uma melhor compreensão dos dados e das propriedades dos algoritmos de ML. As entidades da Figura 5 (ML Schema) podem estar associadas a uma Definição de Ação e são exemplos de atributos armazenáveis. Na figura citada não constam os atributos relativos à gerência de projetos (recursos, prazos, etc.).

Para clarear o conceito, seguem exemplos de definições de ação no projeto hipotético de predição de preço de uma casa[10]:

- Momento de registro: 10/10/2018; Definição da ação: Experimentar como características do imóvel no modelo o número de pavimentos e o número de banheiros.
- Momento de registro: 8/11/2018; Definição da ação: Avaliar se para a característica *distância do centro* o modelo alcança uma melhor performance se os valores forem ajustados para uma outra escala.
- Momento de registro: 1/2/2019; Definição da ação: Experimentar algoritmos *lightgbm* e *randomforest* para a regressão.

3.3 REGISTRO DE TREINAMENTO

Atividade responsável por documentar os atributos envolvidos em um treinamento que podem ser agrupados em 6 categorias[11]:

---

[9] A depender, claro, da existência ou não de requisitos corporativos.

[10] Esses registros podem ser categorizados, entre outras possibilidades, quanto à tarefa CRISP-DM associada. No caso, correspondem, respectivamente, a *Selecionar dados*, *Formatar dados* e *Selecionar técnica*.

[11] A lista não tem a intenção de ser completa, mas ilustrativa.

- Dados usados: dados de teste, dados de treinamento, variáveis usadas, etc.;
- Parâmetros de treinamentos: algoritmo usado, hiperparâmetros considerados, implementações de técnicas aplicadas, etc.;
- Parâmetros de testes: métrica considerada, forma de apuração, etc.;
- Resultados obtidos: modelo, seus parâmetros, métricas apuradas, etc.;
- Dados de configuração: identificação do programa, versões das bibliotecas usadas, dados de hardware, etc.;
- Dados de contexto: código do treinamento, data e hora, número de épocas de treinamento, mensagem de erro em caso interrupção da execução, etc.

A atividade ocorre no estágio de desenvolvimento[12] do processo DM, que engloba, como visto com MARBÁN *et al* (2007), a coleta e a análise dos dados disponíveis para o projeto, a criação de novos dados a partir dos disponíveis, a adaptação para algoritmos de DM e a criação de modelos.

A Figura 5 apresentada no Referencial Teórico traz a ontologia *machine learning schema* (ML Schema) que fornece um conjunto de entidades envolvidas em um treinamento (*W3C Machine Learning Schema Community Group*, 2017) e que são candidatas a terem informações documentadas no contexto da atividade de registro de treinamento.

O treinamento, objeto da atividade, é a atividade central de todo o processo de mineração de dados e é importante obter o máximo possível de informações deles, segundo CHOLLET (2017). Como vimos, o autor detalha que são várias iterações de experimentação de modelos (padrões) que buscam validar uma ideia e cujos resultados inspiram novas ideias, e, quanto mais iterações desse círculo repetitivo forem executadas, mais refinadas e poderosas se tornam as ideias, e, por consequência, os modelos.

É imprescindível que essa atividade seja automática e esteja vinculada à realização de treinamento de modelo na plataforma em uso. Ainda que haja um custo inicial de construção do arcabouço de software para uma determinada configuração de ferramentas, esse esforço será implementado uma única vez e reaproveitado nos demais projetos. GREFF *et al* (2017) alertam sobre o desafio prático da documentação face ao número significativo de experimentos computacionais com inúmeras e diversas configurações de hiperparâmetros quando não se incentiva a construção dessa estrutura, por exemplo, por pressão de prazo.

[12] MARBÁN *et al* (2007) agrupam as tarefas de um processo de DM em estágios de pré-desenvolvimento, desenvolvimento e pós-desenvolvimento em relação à construção de um modelo.

Entre outros usos, o código de um treinamento (no grupo contextual), como será mostrado na experiência do projeto Cladop, pode ser usado como chave de identificação do modelo treinado.

Dada a flexibilidade defendida no Rastro-DM, o nível de detalhe de documentação que impacta a sua utilidade dependerá da equipe e claro da plataforma em uso. Se a preocupação da equipe for com a reprodutibilidade do experimento, por exemplo, há que se gravar mais detalhes de configuração de software (versões de bibliotecas usadas, por exemplo) e de sementes de números aleatórios usados.

Para melhor compreensão dos conceitos aqui definidos, iremos usar um projeto simples hipotético de aprendizagem supervisionada que objetiva a predição (regressão) do preço de uma casa para uma imobiliária a partir de alguns atributos do imóvel.

Abaixo exemplos de registros de treinamentos do projeto hipotético[13]:

- Código: 1; Momento: 7/6/2018; Variáveis usadas: área da casa, área do lote e CEP do endereço; algoritmo usado: *linear regression*; Erro: 0.8; Separação de dados de teste: 10%, não estratificada; Local do modelo gerado: C:\modelos\modelo_cod_1.model;
- Código: 100; Momento: 7/7/2018; Variáveis usadas: área da casa, área do lote, CEP do endereço, número de quartos e data da construção; algoritmo usado: *randomForest*; Erro: 0.7; Separação de dados de teste: 5%, dados estratificados; Local do modelo gerado: C:\modelos\modelo_cod_1.model;

3.4 SÍNTESE DE APRENDIZADO

Atividade responsável por sintetizar e registrar de forma automática, ou não, os aprendizados concebidos ao longo do projeto.

A atividade de síntese de um aprendizado pode ocorrer a qualquer momento de um projeto de DM e pode estar associada ou não a treinamentos, uma vez que aprendizados podem ser sintetizados no estágio de pré-desenvolvimento, nas fases CRISP-DM de entendimento do negócio e dos dados. Mas a maioria dos aprendizados são gerados a partir das experimentações realizadas. Pode-se sintetizar, automaticamente ou não, que uma determinada seleção de variáveis ou que o uso de um determinado hiperparâmetro de uma técnica levou à geração de um modelo de melhor performance. Pode haver aprendizados que envolvem treinamentos que

[13] São exemplos hipotéticos e sem todos os detalhes possíveis.

não conseguiram ser executadas, que objetivam documentar como evitar que um erro aconteça novamente.

Um Aprendizado pode ter como atributo informações das entidades da Figura 5 (ML Schema), mas também pode envolver outros conceitos, como Ações Definidas, Treinamentos e, até Aprendizados[14].

A síntese de aprendizados e o uso efetivo deles no projeto ou em projetos futuros promovem o amadurecimento da equipe no processo de DM, no conhecimento sobre a alta iteratividade e interatividade de suas tarefas, sobre as várias técnicas e ferramentas usadas, etc.

A documentação dos aprendizados impede que eles se percam nas memórias dos indivíduos ou mesmo com os indivíduos quando deixam o projeto ou a organização.

Para clarear o conceito, seguem exemplos de exemplos de possíveis aprendizados concebidos no projeto hipotético de predição de preço de uma casa[15]:

- Em 3/10/2018; Aprendizado: A técnica *randomforest* se mostrou superior à técnica *decisiontree* no contexto avaliado.
- Em 26/5/2019; Aprendizado: É necessário que os valores dos imóveis nos dados usados para treinamento sejam atualizados para uma mesma referência monetária. Criada uma nova coluna de nome *valor_atualizado_em_dezembro_2018*.
- Em 26/7/2019; Aprendizado: O acréscimo das variáveis número de quartos e número de vagas para automóveis promoveu uma melhora de 10% no modelo.

3.5 VISÃO INTEGRADA DOS CONCEITOS DO RASTRO

Os conceitos do rastro se relacionam: os treinamentos acontecem no contexto de uma ação definida no projeto e esses treinamentos podem promover aprendizados que por sua vez podem influenciar, em um círculo virtuoso, novas ideias definidas em ações. A Figura 6[16] apresenta os conceitos envolvidos no Rastro-DM. Percebe-se que o conceito *Atributo de treinamento* pode ser um dos seis grupos citados na seção 3.3 e pode estar envolvido como objetos nas definições de ação. As definições de ação podem ter fundamentação em tarefas da

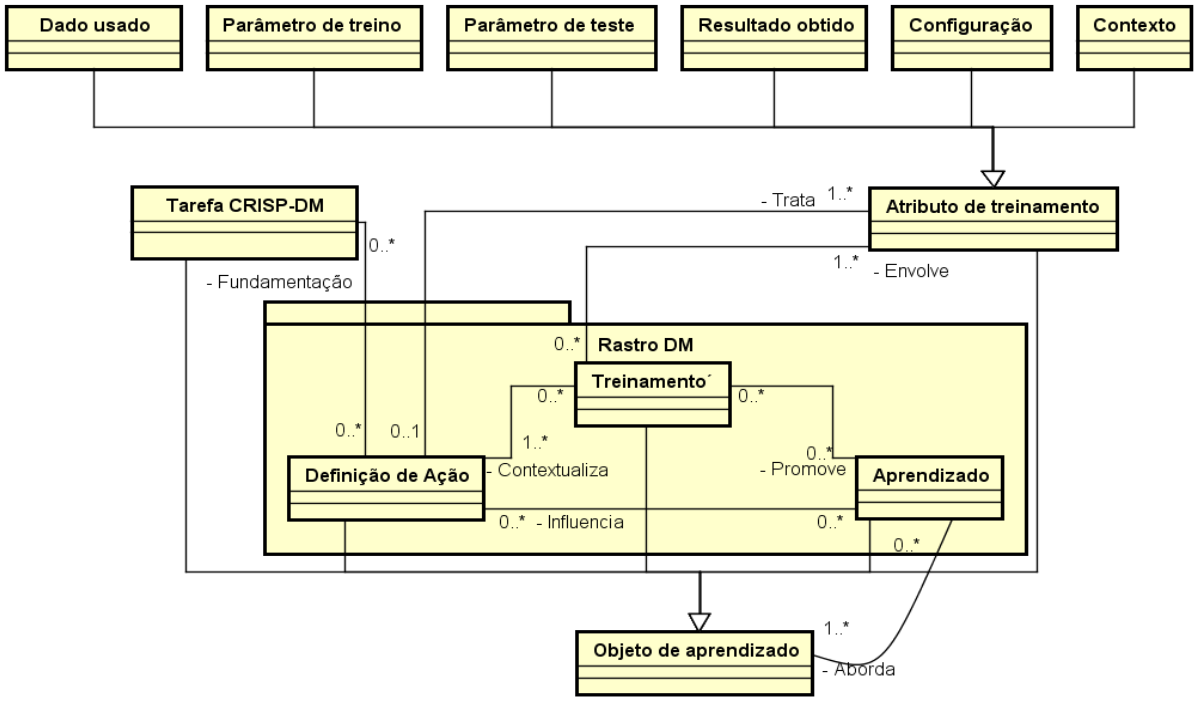
[14] Pode-se, por exemplo, aprender que um aprendizado não faz sentido ou tem um escopo mais limitado.

[15] Esses registros podem ser categorizados, entre outras possibilidades, quanto à tarefa CRISP-DM associada. No caso, correspondem, respectivamente, a: *Selecionar técnica, Formatar dados* e *Selecionar dados*.

[16] A figura não tem a pretensão de ser completa, mas de ilustrar para uma melhor compreensão os conceitos do rastro. Ela não traz, por exemplo, a possibilidade de autorrelacionamento entre as definições de ação. Uma ação definida como *Avaliar o impacto do modelo com variáveis categóricas* pode estar no contexto de outra ação maior *Selecionar variáveis para o modelo*.

metodologia base, representada na figura pelo CRISP-DM. E todos os elementos podem ser abordados em um aprendizado.

Figura 6 – Principais conceitos associados ao Rastro-DM



Fonte: elaborada pelo autor (2019)

A documentação do rastro, fora do contexto rígido de uma *metodologia base*, permite tratar melhor a interatividade e iteratividade das tarefas, algo não bem mapeado pelas metodologias de DM, conforme discutido no Referencial Teórico. A realização das atividades Rastro-DM gera documentação que em um momento posterior, se necessário, pode ser agrupada[17] por tarefa CRISP-DM. E a documentação gerada pela proposta pode se encaixar nos artefatos de saída das *metodologias bases*. Por exemplo, o documento *Razões para exclusões e seleções* da atividade *Seleção de dados* da fase *Preparação de dados* do CRISP-DM pode ser um relatório construído automaticamente a partir dos aprendizados e definições de ações que tratam itens desse contexto e, se desejável, ilustrado com um resumo dos resultados dos treinamentos relacionados a cada critério de seleção de dados experimentado[18].

Em relação às definições de ação e aprendizados, deve-se priorizar o seu registro em detrimento à sua categorização, de forma que esta atividade não seja uma barreira para a

[17] Como as tarefas e as fases de um processo de DM se misturam e muitas vezes são executadas de forma concomitante, fica difícil se manter uma documentação efetiva por tarefa da metodologia base. Se uma tarefa CRISP-DM gera uma documentação e depois volta-se a essa tarefa diversas vezes, a documentação precisaria ser constantemente atualizada de forma restrita ao escopo da tarefa. O custo do retrabalho acaba sendo impeditivo para que a equipe realize a documentação tempestiva.

[18] Trata-se de uma versão inicial do relatório, que poderá ser enriquecida a depender da riqueza de detalhes da documentação.

documentação. Se necessário, pode-se deixar para um passo posterior a sua categorização, adequando-o aos requisitos do projeto ou corporativos, que são importantes pois objetivam aumentar a utilidade da documentação e potencializar a partilha do conhecimento. E, sempre que possível, a categorização deve ser automática. Por exemplo, uma definição de ação pode se dar inicialmente em formato texto livre. Posteriormente, a classificação, por exemplo, quanto às técnicas, hiperparâmetros e atividades envolvidas pode ser realizada[19].

É desejável que as atividades sejam realizadas de forma integrada às plataformas do projeto, para não comprometer o progresso natural do trabalho e o esforço não ser uma barreira para a atividade. Afinal, para ser útil, um modelo de documentação deve corresponder, tanto quanto possível, ao modo como as pessoas trabalham (BECKER e GHEDINI, 2005). O Apêndice A ilustra uma implementação simples de uma infraestrutura de construção de um rastro em *python*. O código não objetiva estar completo ou mesmo sem erros, mas sim ser uma demonstração de que um rastro pode ser criado de forma também simples, a um baixo custo, até mesmo com arquivos locais.

Com o adequado e tempestivo registro do rastro, evita-se o desperdício com execuções repetidas de trabalhos. BECKER e GHEDINI (2005) atribuem o desperdício de se refazer um trabalho ao fato de ser impossível se lembrar, com o transcorrer do projeto, quais treinamentos foram realizados, os conjuntos de dados utilizados, os hiperparâmetros usados e os resultados que foram derivados dos conjuntos de dados.

O rastro, com o amadurecimento da organização em processos de DM, deve se tornar corporativo para a formação de uma base de apoio a projetos. Através de consultas a essa base, pode-se, por exemplo, encontrar projetos que experimentaram determinada técnica ou hiperparâmetros, para o caso de uma equipe desejar mais informações sobre o seu uso.

Rastro-DM foca na documentação do processo por trás da construção do produto final[20] (modelo implantado). Conforme CONKLIN (1996), promove-se assim o aumento da memória organizacional, com o registro do contexto de criação dos artefatos: os pressupostos, os valores, as experiências, o motivo, as conversas e as decisões conduzidas.

[19] Se a tarefa CRISP-DM for um atributo desejável para uma definição de ação e se a descrição for sobre a experimentação de um hiperparâmetro, pode-se automaticamente inferir que a fase é de modelagem. Eis um exemplo de uso de DM para o KMP, combinação que potencializa o crescimento do conhecimento organizacional (NGUYEN, 2018). As organizações com uma padronização mínima na estrutura dos rastros por tipo de tarefa de DM podem facilitar a exploração dessa combinação.

[20] O foco tradicional é na documentação do produto final (e de seus artefatos).

4 PROJETO CLADOP – ESTUDO DE CASO

O projeto Cladop[21] é um estudo de caso da utilização do Rastro-DM. Consistiu no desenvolvimento por aprendizagem supervisionada de um classificador automático de tipo para documentos em formato PDF (*Portable Document Format*) inseridos no sistema de gestão de Tomadas de Contas Especiais (e-TCE) do Tribunal de Contas da União (TCU). Um processo de Tomada de Contas Especial, em última análise, objetiva o ressarcimento do Erário Público de danos gerados por agentes públicos e a devida responsabilização destes. Todos os órgãos da administração pública federal são potenciais usuários do sistema e-TCE.

A seguir apresenta-se a compreensão do contexto e dos dados envolvidos no projeto Cladop e detalhes do classificador gerado. Em seguida descreve-se o rastro do projeto e dois benefícios indiretos do rastro: a integração com uma rotina de monitoramento automático e a geração semi-automática de um relatório.

4.1 COMPREENSÃO DO CONTEXTO DE NEGÓCIO

É competência constitucional do Tribunal de Contas da União (TCU) julgar as contas daqueles que derem causa à perda, extravio ou outra irregularidade, com dano, prejuízo, ao Erário, conforme artigo 71, inciso II da Constituição Federal (BRASIL, 1988).

Esse julgamento se dá através de um processo administrativo devidamente formalizado, com rito próprio para apuração de responsabilidade e obtenção de ressarcimento por ocorrência de dano à administração pública federal, com apuração de fatos, quantificação do dano e identificação dos responsáveis, pessoas físicas ou jurídicas às quais possa ser imputada a obrigação de ressarcir o Erário. Esse processo recebe o nome de Tomada de Contas Especial (TCE), segundo o artigo 2º da Instrução Normativa 71/2012 do TCU (BRASIL, 2012).

Conforme o artigo 5º da referida IN, é pressuposto para instauração de tomada de contas especial a existência de elementos fáticos e jurídicos que precisam ser lastreados em documentos, narrativas e outros elementos probatórios que deem suporte à sua ocorrência. O parágrafo primeiro do artigo décimo complementa que o relatório do tomador de contas deve informar a localização nos autos sempre que mencionar esses documentos, e os divide em quatro grupos: (a) documentos utilizados para demonstração da ocorrência de dano; (b) notificações remetidas aos responsáveis, acompanhadas dos respectivos avisos de recebimento

[21] Cladop é também uma referência ao modelo gerado: Classificador de Documentos em PDF.

ou de qualquer outro documento que demonstre a ciência dos responsáveis; (c) pareceres emitidos pelas áreas técnicas do órgão ou entidade, incluída a análise das justificativas apresentadas pelos responsáveis; (d) e outros documentos considerados necessários ao melhor julgamento da tomada de contas especial pelo TCU. A importância dos documentos é tal que o artigo 13 da referida IN estabelece que, se faltar alguma peça necessária, o processo deve ser devolvido pelo TCU ao órgão de controle interno.

A Decisão Normativa 155/2016 do TCU (BRASIL, 2016) vai mais além e estabelece quais documentos, ao indicar seus tipos, devem ser enviados juntos ao relatório do tomador de contas. Detalha os documentos para os grupos (a) e (d), do §1º do art. 10 da IN citado antes, e ainda relaciona documentos que precisam ser encaminhados a depender da origem dos recursos cujo desvio será apurado. Como exemplo de documentos por grupo, estabelece que devem ser encaminhados como comprovação de ocorrência de dados (grupo a): ordens bancárias, ou equivalente que demonstre a execução financeira, e notas de empenho, ou equivalente que demonstre a execução orçamentária. Como exemplo de documentos obrigatórios por origem de recursos, estabelece que, se os recursos forem transferidos por meio de termo de compromisso com o CNPq e Capes, deve ser juntada uma cópia do documento do termo de concessão e de aceitação da bolsa e aditivos.

A referida DN, em seu anexo 1, sugere medidas administrativas para auxiliar, em caráter subsidiário e facultativo, o órgão ou entidade instauradora da TCE, na adoção das medidas administrativas, com vistas à apuração dos fatos, identificação dos responsáveis e obtenção do ressarcimento do dano. Sugere, por exemplo, formato padronizado de diligência que deve ser realizada com vistas à obtenção de informações ou de documentos necessários à elucidação dos fatos.

Com o objetivo de tornar mais célere e eficaz o trâmite do processo de apuração de danos, com padronização e otimização de procedimentos, foi desenvolvido pelo TCU com cooperação da CGU o sistema e-TCE, plataforma única de acesso a todas as entidades da Administração Pública que atuam em alguma fase da TCE.

A Portaria do TCU 122/2018 (BRASIL, 2018), que dispõe sobre a implantação do sistema e-TCE, estabelece em seu artigo 11 que o sistema deve conter uma lista de tipos de documentos que contemple os documentos relacionados na DN 155/2016. Afirma ainda que devem ser inseridos de acordo com a ordem cronológica constante no processo administrativo originário. Mas deixa aberto em seu parágrafo primeiro que outros tipos deverão ser incluídos no Sistema sempre que necessários à demonstração da ocorrência de dano ou melhor apreciação

do processo. Ou seja, deixa em aberto a relação de tipos, o que justifica a aceitação pelo sistema do tipo *outros*, que deve ser detalhado por uma descrição complementar.

O parágrafo segundo do mesmo artigo estabelece que a ausência dos documentos obrigatórios e de outras peças que fundamentem o relatório do tomador de contas deverá ser objeto de justificativa, embasada, quando for o caso, em elementos que demonstrem as tentativas de obtenção da referida documentação. Mas, conforme o artigo 17 da citada Portaria, o TCU poderá devolver a TCE ao órgão do sistema de controle interno, antes da autuação, caso entenda necessária a realização de ajustes e a complementação de informações.

Segundo o artigo 19 da mesma Portaria, o Sistema e-TCE funciona como protocolo eletrônico dos órgãos instauradores, de controle interno, da autoridade supervisora e do TCU para efeito de tramitação de TCE e de documentos complementares e de envio e atendimento de comunicações processuais. Conforme o artigo 31, cada entidade deve adotar medidas de segurança e salvaguarda dos documentos originais que compõem a TCE, com vistas a preservar a integridade e a autenticidade de documentos e de dados inseridos no Sistema e-TCE que são considerados originais para todos os efeitos até prova em contrário e devem observar o formato PDF e as especificações disponíveis nos tutoriais do sistema (artigo 16).

Nesse contexto, o classificador automático de tipo de documento, Cladop, desenvolvido no projeto, promoveu benefícios diretos e indiretos ao negócio.

Como benefício direto, ele permite a correção dos tipos dos documentos inseridos, evitando-se erros de classificação: ao identificar, automaticamente, qual o tipo de documento mais adequado, no momento de inserção de cada documento comprobatório; ao evitar que sejam classificados como *outros* documentos de tipos previstos, evitando que usuários por comodidade optem por selecionar essa categoria coringa; ao poder ser usado para corrigir a classificação de documentos indevidamente catalogados como *outros*.

Como benefício indireto, ele pode apoiar a promoção de uma melhor qualidade do OCR dos documentos: ao levar em conta um conteúdo *esperado mínimo* para se classificar um documento com maior probabilidade de acerto; ao subsidiar a construção de críticas no sistema de conteúdo mínimo por tipo, retornando, além das previsões de tipo, informações complementares resultantes do pré-processamento do texto dos documentos, como contagens de palavras por classe (nome, UF, CPF/CNPJ, data e número) e indicadores de qualidade do PDF (quantidade de palavras válidas, palavras desconhecidas, páginas, etc.). De posse dessas informações, o Sistema e-TCE pode, por exemplo, impedir que um documento do tipo AR (*Aviso de Recebimento*) seja inserido se não houver pelo menos um CPF ou CNPJ ou mesmo impedir documentos com 50% de palavras inválidas, evitando-se, assim, uma baixa qualidade

para tipos específicos e críticos para o processo de instrução. Afinal, documentos com classificação correta de tipo e com um conteúdo OCR de melhor qualidade são fundamentais para os passos seguintes do rito processual, como a instrução assistida pelo computador.

4.2 ENTENDIMENTO DOS DADOS

Há um acentuado desbalanceamento na quantidade de documentos por tipo e o tipo coringa *Outros* é o mais usado com 18,81% dos documentos, ou seja, 22.253 de 118.266. A Tabela 1 discrimina o quantitativo por tipo e o Gráfico 1 permite a visualização do desbalanceamento.

Tabela 1 - Total de documentos por tipo (referência: 17/4/2019)

Tipo documento	Documentos
Acórdão	145
Análise de Prestação de Contas	502
Análise defesa	335
Análise e Avaliação técnica do Relatório Final	108
Ata de aprovação do projeto pelo Conselho/Comissão	412
Ata/portaria/decreto de nomeação e exoneração	1.401
Avaliação da execução do projeto	120
Avaliação dos objetivos e finalidades da instituição-art. 35, III, Lei 13.019/2014	4
Aviso de recebimento (AR) ou equivalente	8.691
Ação judicial - petição inicial	444
Baixa de responsabilidade em apuração	259
Cheque, comprovante de transferência bancária ou outro comprovante de pagamento	1.324
Comprovante de endereço	199
Comprovante de pagamento efetuado ao beneficiário	292
Comprovante de recolhimento de saldo de recursos	835
Conciliação bancária	618
Contrato firmado com a empresa contratada para a exec. da obra ou serviço	738
Cópia do diploma ou declaração de conclusão	14
Declaração de gratuidade	79
Declaração de realização dos objetivos a que se propunha o instrumento	534
Declarações da autoridade local atestando a realização do objeto do convênio	130
Defesa/manifestação do responsável	6.031
Demonstrativo da situação atual das contas - SiGPC e/ou SIAFI	769
Demonstrativo de débito	2.275
Demonstrativo de recursos aprovados e captados	181
Despacho de expediente	3.530
Despacho do controle interno	690
Determinação/recomendação de instauração	2.393
Documento de atesto do recebimento da obra ou serviço	85
Edital de chamamento público	139
Extrato bancário conta específica, da data dos créditos até o encerramento da movimentação	3.189
Ficha de qualificação do responsável	2.768
Fotografia	275

Fotografia do objeto	79
Instrumento que formalizou a parceria e respectivos termos aditivos	3
Instrumento que formalizou a transferência/parceria e respectivos termos aditivos	2.242
Material de divulgação para fins de prest. de contas	598
Matriz de responsabilização	1.193
Normativo que disciplina a concessão da bolsa/auxílio	30
Nota de empenho, ou equivalente que demonstre a execução orçamentária	1.278
Nota de lançamento	2
Nota fiscal ou outro comprovante de despesa	3.530
Nota técnica	1.603
Notificação (ofício), inclusive edital	12.972
Ordem bancária, ou equivalente que demonstre a execução financeira	2.239
Outros	22.253
Parecer com recomendação para aprovação/reprovação do projeto	493
Parecer do órgão técnico da adm. pública - art. 35, V, Lei 13.019/2014	51
Parecer emitido s/exec. física do objeto e do atend. aos objetivos avença	1.146
Parecer financeiro	1.918
Parecer jurídico	872
Parecer jurídico sobre a minuta do instrumento que formalizou a transferência	1.095
Parecer jurídico sobre possibilidade de celeb. da parceria - art. 35, VI, Lei 13.019/2014	11
Parecer téc. e financ. avaliaç. do plano de trabalho	1.103
Parecer técnico/nota técnica/nota explicativa	2.220
Plano de trabalho aprovado	1.566
Portaria de aprovação do projeto	418
Portaria/Despacho inicial de instauração da TCE	2.100
Publicação do extrato do instrumento no D.O.U	1.330
Recibo de incentivo	592
Registro da inadimplência	1.668
Registro de Responsabilidade em Apuração do débito	561
Registro do débito apurado em conta do ativo (Diversos Responsáveis)	3.234
Relatório Físico	250
Relatório de cumprimento do objeto	811
Relatório de execução da receita e da despesa	554
Relatório de execução físico-financeira	685
Relatório de fiscalização do TCU	75
Relatório de fiscalização do órgão de controle interno	382
Relatório de fiscalização do órgão ou entidade repassador	1.411
Relatório de sindicância, inquérito, PAD ou equivalente	239
Relatório de visita técnica in loco	615
Relatório final	565
Relatório parcial	168
Relatório técnico de monitoramento e avaliação da parceria	323
Relação de bens de capital ou de bens imóveis	282
Relação de bens, de serviços prestados ou de treinados/capacitados	755
Relação de pagamentos	1.124
Solicitação de apoio a projetos	490
Suspensão de inadimplência	480
Termo de aprovação/reprovação de prestação de contas	166
Termo de concessão e de aceitação da bolsa e aditivos	135
Termo de recebimento definitivo da obra	68
Termos de homologação e de adjudicação do processo licitatório	779

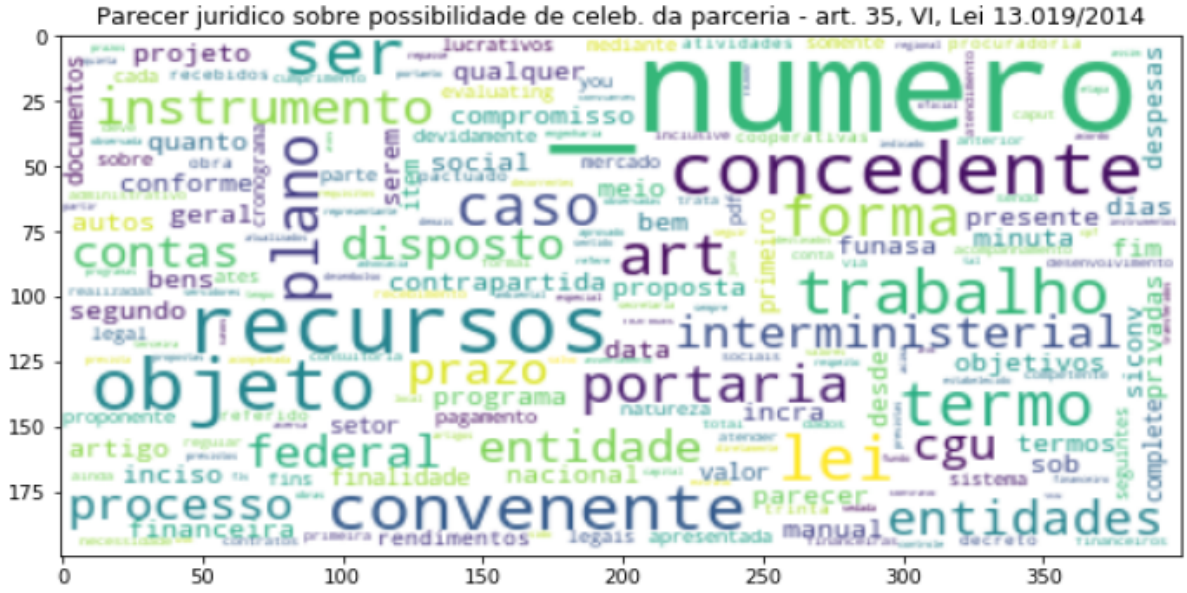
Fonte: elaborada pelo autor (2019)

Gráfico 1 - Total de documentos por tipo (referência: 17/4/2019)

Fonte: elaborada pelo autor (2019)

Para fins de entendimento melhor dos dados, das palavras que caracterizam cada tipo de documento, foram gerados gráficos com nuvens de palavras (consideradas válidas) a partir dos documentos de cada tipo. A Figura 7 ilustra as palavras de documentos do tipo *Parecer jurídico sobre possibilidade de celeb. da parceria - art. 35, VI, Lei 13.019/2014*, de código 26.

Figura 7 - Nuvem de palavras[22] dos documentos do tipo *Parecer jurídico (art. 35)*



Fonte: elaborada pelo autor (2019)

Com o objetivo de exploração e compreensão dos dados, foram levantados alguns conjuntos frequentes associados aos tipos de documentos. A Tabela 2 mostra os conjuntos mais frequentes de termos associados ao tipo *Notificação (ofício), inclusive edital*, de código 42. Para esse tipo, foram identificados 1283 conjuntos frequentes para um suporte mínimo indicado de 40%, isto é, conjuntos presentes em pelo menos 40% dos documentos. A quantidade de agrupamentos por número de palavras do conjunto segue a seguinte distribuição: {1: 34, 2: 202, 3: 437, 4: 404, 5: 168, 6: 35, 7: 3}.

[22] Nuvem de palavras é uma imagem com várias palavras em que o tamanho da fonte de cada termo é proporcional à frequência da palavra nos textos. Quanto mais frequente a palavra, maior o tamanho de sua fonte. Na Figura 7, *_numero* é o termo mais comum. Antes da construção da figura, foi realizado um pré-processamento com o texto dos arquivos que substituiu os números pelo termo *_numero* e trocaram-se maiúsculas por minúsculas.

Tabela 2 - Conjuntos mais frequentes do tipo *Notificação (ofício), inclusive edital*

Suporte	Conjuntos
1.00000	(valor)
0.69046	(municipal, valor)
0.69046	(municipal)
0.62769	(valor, prefeitura)
0.62769	(prefeitura)
0.62425	(iii)
0.62425	(iii, valor)
0.60748	(estado, valor)
0.60748	(estado)
0.59802	(contas, valor)

Fonte: elaborada pelo autor (2019)

Também foram identificadas regras de associação entre os conjuntos frequentes[23]. A Tabela 3 apresenta as principais regras para o tipo *Notificação (ofício), inclusive edital* (código 42). Foram selecionadas 6891 regras considerando um percentil de corte de 75% e um *lift* mínimo de 1.2.

Tabela 3 - Principais regras de associação para o tipo *Notificação (ofício), inclusive edital*

Antecedentes	consequentes	support	confidence	Lift	leverage	conviction
(ser)	(cpf)	0.47635	0.92487	1.74899	0.20399	6.27213
(cpf)	(ser)	0.47635	0.90081	1.74899	0.20399	4.88928
(ser, valor)	(cpf)	0.47635	0.92487	1.74899	0.20399	6.27213
(cpf, valor)	(ser)	0.47635	0.90081	1.74899	0.20399	4.88928
(ser)	(cpf, valor)	0.47635	0.92487	1.74899	0.20399	6.27213
(cpf)	(ser, valor)	0.47635	0.90081	1.74899	0.20399	4.88928
(processo)	(secretaria)	0.44153	0.80047	1.69725	0.18139	2.64807
(secretaria)	(processo)	0.44153	0.93619	1.69725	0.18139	7.02721
(processo, valor)	(secretaria)	0.44153	0.80047	1.69725	0.18139	2.64807

Fonte: elaborada pelo autor (2019)

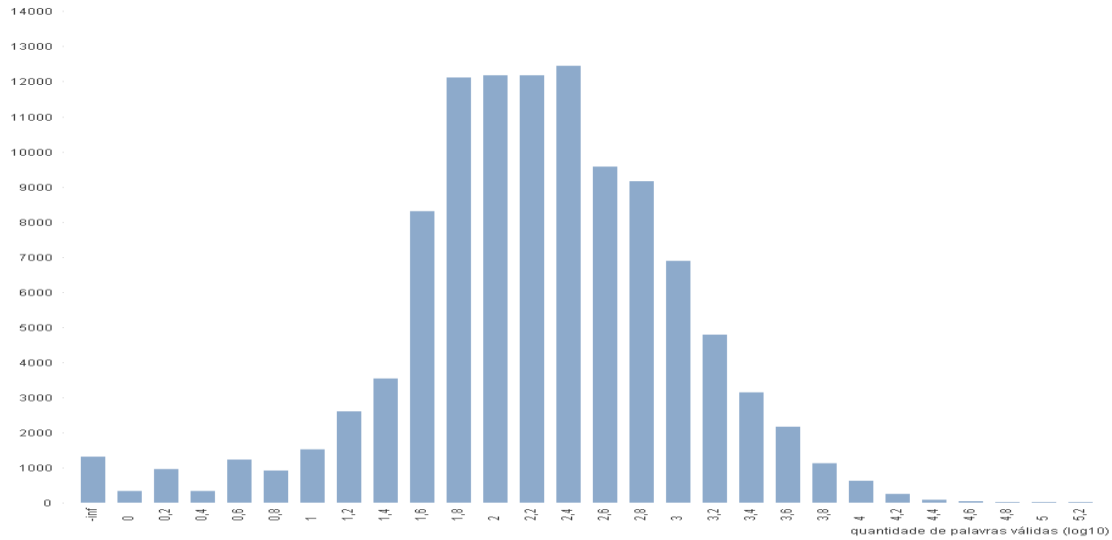
BRANTING (2017) afirma que o formato PDF tem sido usado em tribunais e que o texto obtido desses documentos apresenta muitos erros e não preservam a sequência original do documento, devido ao processo usado de OCR (*Optical Character Recognition*).

O alcance dos resultados esperados pelo sistema e-TCE depende da qualidade dos documentos protocolados no sistema. A eficácia do classificador também tende a ser superior se os dados tiverem uma maior qualidade. Foi encontrada uma baixa qualidade no OCR dos documentos. Um exemplo dessa situação é o documento de protocolo 58.900.414 que tem 162 páginas, mas não se consegue via OCR identificar nem uma centena de palavras válidas, ou seja, menos de uma palavra por página. A qualidade do OCR pode ser ilustrada pelos

[23] Regras de associação são usadas para identificação de padrões de coexistência entre conjuntos frequentes em um documento. Mais detalhes em https://pt.wikipedia.org/wiki/Regras_de_associação (acesso em 31/8/2019)..

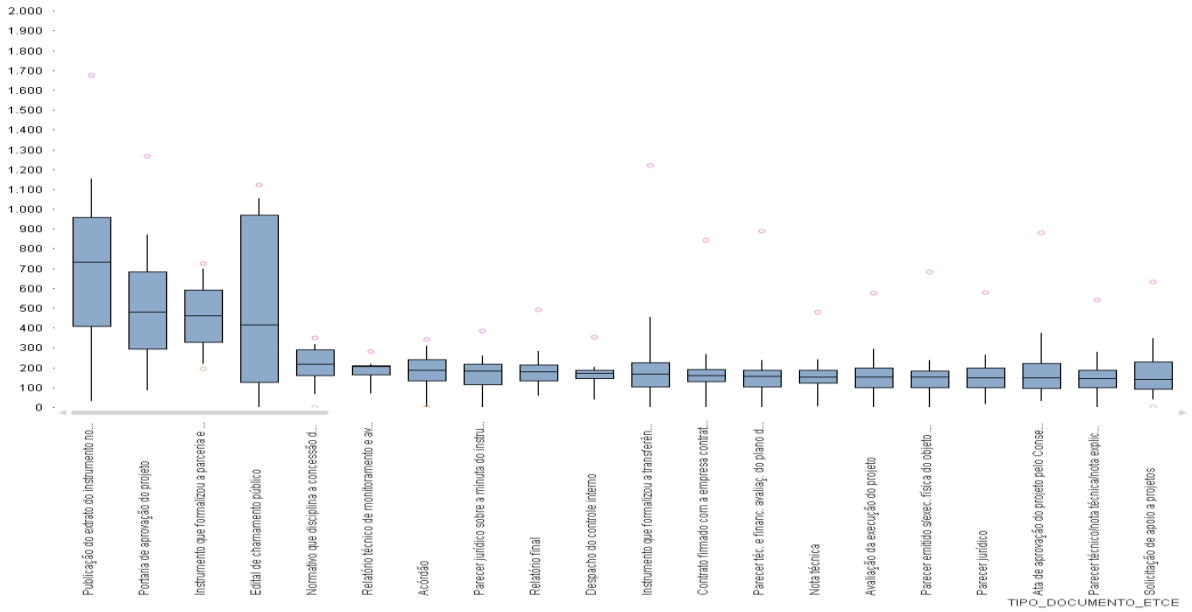
gráficos que se seguem: o Gráfico 2 ilustra um histograma[24] do atributo quantidade de palavras válidas e o Gráfico 3 mostra um *boxplot*[25] da distribuição de palavras válidas por página para alguns tipos de documento.

Gráfico 2 - Histograma de quantidade de palavras válidas (log10).



Fonte: elaborada pelo autor (2019)

Gráfico 3 - *Boxplot* da distribuição de palavras válidas por página para alguns tipos.



Fonte: elaborada pelo autor (2019)

[24] Conforme a *wikipedia*, histograma, também conhecido como distribuição de frequências, é a representação gráfica em colunas em que a base de cada retângulo representa uma classe e a altura representa a quantidade ou a frequência absoluta com que o valor da classe ocorre no conjunto de dados.

[25] Conforme a *wikipedia*, *boxplot*, também conhecido como diagrama de caixa, representa a variação de dados por meio de quartis. A caixa é dividida em 2 partes pela mediana: quartil superior e inferior. Há uma reta que se estende a partir da caixa e indica a variabilidade fora desses quartis. Os valores atípicos ou *outliers* podem ser plotados como pontos individuais.

Os dados do sistema e-TCE foram replicados em uma base espelho para que o projeto não concorresse nos acessos aos dados com os usuários do sistema. E também para que tivesse uma certa autonomia estrutural, uma vez que essa base não segue as normalizações necessárias em um sistema on-line.

Além dos atributos dos Danos e de seus documentos, a base armazena para cada documento o texto após o pré-processamento do conteúdo do arquivo e alguns atributos de qualidade do OCR. Detalhes dessa base constam do Apêndice B.

Foram experimentadas formas diferentes de pré-processamento do texto para extração das palavras válidas dos conteúdos dos arquivos. Durante a evolução do projeto, métodos novos foram substituindo os anteriores. Os seis métodos experimentados constam da Tabela 4 e retratam diferentes combinações dos seguintes passos:

- Derivação de 7 classes substitutas de palavras (cpf/cnpj, números, datas, nomes de pessoas físicas, nomes e siglas de estados). Exemplificando, se no texto há o nome rio de janeiro, faz-se a substituição dessas 3 palavras por uma única palavra também considerada válida *classenomeuf*. O último método (com identificação 7) usa todas as classes.

- Validação se um *token* é uma palavra correta se a mesma existir em um conjunto de palavras válidas. Conjuntos diferentes foram experimentados: palavras de acórdãos do TCU, palavras selecionadas da wiki em português e palavras selecionadas da *wiki* com acréscimo de abreviações e siglas típicas do negócio. O método atual usa o último conjunto, mais adequado ao negócio.

- Obtenção do texto do PDF, seja por OCR original ou por OCR extra com tesseract. Durante a evolução do projeto, dada a necessidade de um curto tempo de resposta do classificador para retornar uma previsão de tipo, abandonou-se a execução de OCR extra com tesseract, que demorava alguns minutos por documento.

Tabela 4 - Métodos de pré-processamento usados no projeto

Código	Descrição
2	Conteúdo novo OCR com Tesseract 4.0; dicionário com palavras de acórdãos do TCU; 2 classes número e cpf/cnpj. palavras transformadas para minúsculas.
3	Conteúdo original com Pdftotext; dicionário com palavras de acórdãos do TCU; 2 classes número e cpf/cnpj; palavras transformadas para minúsculas.
4	Conteúdo novo OCR com Tesseract 4.0; dicionário com palavras selecionadas wiki; 2 classes número e cpf/cnpj; palavras transformadas para minúsculas.
5	Conteúdo OCR original com Pdftotext; dicionário com palavras selecionadas wiki; 2 classes número e cpf/cnpj; palavras transformadas para minúsculas.
6	Conteúdo OCR original com Pdftotext; dicionário com palavras selecionadas wiki após filtro; 2 classes número e cpf/cnpj; palavras transformadas para minúsculas.
7	Conteúdo OCR original com Pdftotext; dicionário com palavras selecionadas wiki e termos comuns do negócio; 7 classes; palavras transformadas para minúsculas.

Fonte: Elaborada pelo autor (2019).

4.3 DESCRIÇÃO FUNCIONAL DO CLASSIFICADOR

O classificador, em sua versão 2.1, alcança a acurácia[26] de 91,1% com desvio padrão de 0,3%, foi implementado em *python*[27] na forma de um *webservice* que recebe como parâmetro um arquivo PDF, e a partir do seu nome e do seu conteúdo retorna nove[28] tipos mais prováveis, com suas respectivas probabilidades.

Tabela 5 mostra algumas métricas por tipo de documento apuradas quando da geração da versão em produção do classificador[29].

Tabela 5 - Métricas apuradas por tipo sobre os dados de validação para a versão em produção do Cladop.

Tipo	Descrição	Precisão	Recall	F1	Documentos
1	Ordem bancária, ou equivalente que demonstre a execução financeira	98,46%	98,46%	98,46%	130
2	Nota de empenho, ou equivalente que demonstre a execução orçamentária	98,61%	94,67%	96,60%	75
3	Relação de pagamentos	86,00%	89,58%	87,76%	48
4	Relatório de execução físico-financeira	88,24%	83,33%	85,71%	36
5	Relatório de cumprimento do objeto	95,12%	81,25%	87,64%	48
6	Declaração de realização dos objetivos a que se propunha o instrumento	91,18%	91,18%	91,18%	34
7	Relação de bens, de serviços prestados ou de treinados/capacitados	78,95%	90,91%	84,51%	33
8	Extrato bancário conta espec., da data dos créd. até o encer. movimentação	96,89%	98,73%	97,81%	158
9	Nota fiscal ou outro comprovante de despesa	97,67%	96,00%	96,83%	175
10	Documento de atesto do recebimento da obra ou serviço	100,00%	100,00%	100,00%	1
11	Parecer téc. e financ. avaliaç. do plano de trabalho	72,73%	78,43%	75,47%	51
13	Parecer jurídico sobre a minuta do instrumento que formalizou a transferência	70,91%	84,78%	77,23%	46
14	Instrumento que formalizou a transferência/parceria e respectivos termos aditivos	91,20%	91,94%	91,57%	124
15	Parecer emitido s/exec. física do objeto e do atend. aos objetivos avença	65,46%	65,46%	65,46%	55
16	Comprovante de recolhimento de saldo de recursos	94,87%	94,87%	94,87%	39
17	Cheque, comprovante de transferência bancária ou outro comprovante de pagamento	88,24%	81,82%	84,91%	55
18	Relatório de fiscalização do órgão ou entidade repassador	88,75%	92,21%	90,45%	77

[26] Acurácia apurada com validação cruzada de 7 partições e 14 amostras. No contexto deste trabalho, sempre que referenciarmos o termo acurácia, se nada for dito em contrário, deve-se subentender acurácia micro, que leva em consideração os resultados, acertos e erros, por documento independentemente do seu tipo.

[27] Detalhes técnicos da construção do modelo serão apresentados na Seção 4.4.2.

[28] Foi constatado que a acurácia do classificador sobe de 91% (do primeiro tipo) para 99% quando consideradas as nove primeiras previsões. Possibilita que o sistema, por exemplo, apresente os nove tipos em uma segunda tela caso o primeiro tipo não seja aprovado pelo usuário.

[29] Métricas apuradas sobre os dados de validação, 5% do total. Números gerais: Acurácia: 90.99; Precisão: {macro: 81,04%, *weighted*: 91,06%}; *Recall*: {macro: 80,29%, *weighted*: 90,99%}; F1: {macro: 80,16%, *weighted*: 90,87%}

19	Relatório de fiscalização do órgão de controle interno	64,29%	90,00%	75,00%	10
20	Contrato firmado com a empresa contratada para a exec. da obra ou serviço	91,49%	97,73%	94,51%	44
21	Termo de recebimento definitivo da obra	100,00%	100,00%	100,00%	4
22	Termos de homologação e de adjudicação do processo licitatório	94,29%	94,29%	94,29%	35
23	Parecer do órgão técnico da adm. pública - art. 35, V, Lei 13.019/2014	50,00%	33,33%	40,00%	3
24	Plano de trabalho aprovado	91,89%	95,78%	93,79%	71
25	Avaliação dos objetivos e finalidades da instituição- art. 35, III, Lei 13.019/2014	0,00%	0,00%	0,00%	0
26	Parecer jurídico sobre possibilidade de celeb. da parceria - art. 35, VI, Lei 13.019/2014	0,00%	0,00%	0,00%	1
28	Relatório técnico de monitoramento e avaliação da parceria	85,00%	85,00%	85,00%	20
29	Termo de concessão e de aceitação da bolsa e aditivos	75,00%	54,55%	63,16%	11
30	Comprovante de pagamento efetuado ao beneficiário	94,44%	85,00%	89,47%	20
31	Cópia do diploma ou declaração de conclusão	0,00%	0,00%	0,00%	1
32	Parecer jurídico	83,72%	70,59%	76,60%	51
33	Relatório final	85,00%	89,47%	87,18%	19
34	Demonstrativo de recursos aprovados e captados	100,00%	100,00%	100,00%	10
35	Relatório de execução da receita e da despesa	93,10%	96,43%	94,74%	28
36	Conciliação bancária	94,60%	92,11%	93,33%	38
37	Parecer técnico/nota técnica/nota explicativa	80,36%	69,77%	74,69%	129
38	Relatório parcial	50,00%	42,86%	46,15%	7
39	Matriz de responsabilização	100,00%	100,00%	100,00%	63
40	Relatório de sindicância, inquérito, PAD ou equivalente	85,71%	85,71%	85,71%	7
42	Notificação, inclusive edital	96,56%	97,11%	96,84%	693
43	Ficha de qualificação do responsável	99,25%	99,25%	99,25%	133
44	Demonstrativo de débito	98,97%	97,96%	98,46%	98
45	Defesa/manifestação do responsável	96,97%	94,12%	95,52%	306
47	Fotografia do objeto	66,67%	40,00%	50,00%	5
48	Fotografia	91,67%	91,67%	91,67%	12
49	Despacho do controle interno	90,00%	79,41%	84,38%	34
50	Determinação/recomendação de instauração	87,88%	79,09%	83,25%	110
51	Despacho de expediente	88,65%	91,62%	90,11%	179
52	Ação judicial - petição inicial	93,33%	93,33%	93,33%	15
53	Análise defesa	83,33%	52,63%	64,52%	19
54	Portaria/Despacho inicial de instauração da TCE	83,19%	92,16%	87,44%	102
55	Publicação do extrato do instrumento no D.O.U	88,33%	94,64%	91,38%	56
56	Registro da inadimplência	95,00%	88,37%	91,57%	86
57	Registro do débito apurado em conta do ativo (Diversos Responsáveis)	91,61%	94,67%	93,12%	150
58	Aviso de recebimento - AR	97,85%	97,64%	97,74%	466
59	Parecer financeiro	75,76%	88,24%	81,52%	85
60	Relatório de fiscalização do TCU	66,67%	100,00%	80,00%	2
61	Edital de chamamento público	77,78%	77,78%	77,78%	9
62	Relatório de visita técnica in loco	58,07%	78,26%	66,67%	23
63	Ata/portaria/decreto de nomeação e exoneração	93,10%	91,53%	92,31%	59
64	Comprovante de endereço	90,00%	81,82%	85,71%	11
65	Demonstrativo da situação atual das contas - SiGPC e/ou SIAFI	86,67%	95,12%	90,70%	41

66	Termo de aprovação/reprovação de prestação de contas	40,00%	40,00%	40,00%	10
67	Análise de Prestação de Contas	77,27%	62,96%	69,39%	27
68	Análise e Avaliação técnica do Relatório Final	0,00%	0,00%	0,00%	5
69	Avaliação da execução do projeto	50,00%	50,00%	50,00%	6
70	Material de divulgação para fins de prest. de contas	80,00%	82,76%	81,36%	29
71	Nota técnica	73,63%	83,75%	78,36%	80
72	Portaria de aprovação do projeto	83,33%	100,00%	90,91%	15
73	Recibo de incentivo	100,00%	100,00%	100,00%	30
74	Relação de bens de capital ou de bens imóveis	92,86%	100,00%	96,30%	13
75	Relatório Físico	100,00%	90,91%	95,24%	11
76	Solicitação de apoio a projetos	75,00%	94,74%	83,72%	19
77	Suspensão de inadimplência	73,68%	82,35%	77,78%	17
78	Ata de aprovação do projeto pelo Conselho/Comissão	81,25%	92,86%	86,67%	14
79	Parecer com recomendação para aprovação/reprovação do projeto	64,29%	39,13%	48,65%	23
80	Acórdão	85,71%	85,71%	85,71%	7
81	Declarações da autoridade local atestando a realização do objeto do convênio	87,50%	100,00%	93,33%	7
82	Declaração de gratuidade	100,00%	100,00%	100,00%	3
83	Registro de Responsabilidade em Apuração do débito	88,89%	64,00%	74,42%	25
84	Baixa de responsabilidade em apuração	91,67%	91,67%	91,67%	12
85	Normativo que disciplina a concessão da bolsa/auxílio	100,00%	50,00%	66,67%	2

Fonte: elaborada pelo autor (2019)

Adicionalmente às predições, o Cladop retorna informações derivadas do pré-processamento do texto do arquivo, que podem ser úteis para o sistema e-TCE exigir no cadastro dos documentos uma qualidade mínima de conteúdo do OCR, como: quantidade de palavras válidas, quantidade de valores, quantidade de nomes, etc. Assim, o sistema pode impedir que documentos críticos para o processo de TCE tenham qualidade baixa de OCR com um alto percentual de palavras inválidas ou mesmo que tenham conteúdo incompleto, como o caso de um documento de AR, Aviso de Recebimento, sem informação de CPF ou CNPJ e data.

A seguir, é apresentada a estrutura do JSON (*JavaScript Object Notation*) retornada pelo *webservice* implantado, contendo para cada chave na estrutura um exemplo de valor e um comentário.

{"**predicoes**": # nove predições com maior probabilidade.
 [
  {
   "**num**": *1*, *# número da previsão. O tipo mais provável tem num == 1*
   "**cod_tipo**": *4*, *# código do tipo previsto*
   "**probabilidade**": *0.3993 # probabilidade da previsão*

 # as informações que se seguem se referem às métricas apuradas para o tipo durante o treinamento da versão final do modelo. Objetiva indicar o valor da métrica apurado na base de teste

"**precisao_tipo_teste**": *0.8824*, *# métrica precisão apurada em teste para o tipo*

"**recall_tipo_teste**": *0.8333*, *# métrica recall apurada em teste para o tipo*

"**f1_score_tipo_teste**": *0.8571*, *# métrica F1 apurada em teste para o tipo*

"**qtd_registro_tipo_teste**": *36*, *# quantidade de registros com o tipo na apuração*

}

],

 # as informações que se seguem se referem a indicadores de qualidade do OCR do documento

"**qualidade_pdf**": {

"**qtd_palavra_dicionario**": *7*, *# quantidade de palavras do texto que são consideradas válidas, que constam do dicionário*

"**qtd_numero**": *0*, *# quantidade de números no texto*

"**qtd_num_cpf_cnpj**": *0*, *# quantidade de números no formato de cpf ou de cnpj*

"**qtd_char**": *74*, *# quantidade de caracteres*

"**qtd_char_numero**": *0*, *# quantidade de caracteres que são números*

"**qtd_char_consoante**": *32*, *# quantidade de caracteres que são consoantes*

"**qtd_char_vogal**": *33*, *# quantidade de caracteres que são vogais*

"**qtd_char_espaco**": *9*, *# quantidade de caracteres que são espaços*

"**qtd_char_pontuacao**": *0*, *# quantidade de caracteres que são pontuação. São considerados pontuação: !?,.;*

"**qtd_char_separador**": *0*, *# quantidade de caracteres que são separadores. São considerados separadores: []{}”’: ()/<>\-|_-*

"**qtd_char_simbolo**": *0*, *# quantidade de caracteres que são símbolos. São considerados símbolos-> #$%&§ªºº*+=@*

"**qtd_token**": *10*, *# quantidade de tokens (sequências com mais de um caracter, letra ou número)*

"**qtd_token_palavra**": *10*, *# quantidade de tokens que não são só números*

"**qtd_trio_consoante**": *2*, *# quantidade de sequências de 3 letras que são consoantes. Em princípio, indica um erro de grafia!*

"**qtd_trio_char_repetido_nesp**": *0*, *# quantidade de sequências de letras repetidas diferentes de espaço. Em princípio, indica um erro de grafia!*

"**qtd_trio_outro**": *0*, *# quantidade de sequências de 3 caracteres que não sejam letras nem espaço. Em princípio, indica um erro de grafia!*

"**qtd_pagina**": *1*, *# quantidade de páginas no documento*

"**qtd_linha**": *1*, *# quantidade de linhas no texto*

"**qtd_char_dif**": *0*, *# quantidade de caracteres não normais (letras, símbolos, separadores e pontuação)*

"**tamanho_token_media**": *6.5*, *# tamanho médio dos tokens*

"**tamanho_token_std**": *2.202271554554524*, *# desvio padrão do tamanho dos tokens*

"**qtd_data**": *0*, *# quantidade de datas*

"**qtd_parte_nome**": *0*, *# quantidade de pedaços de nomes: sobrenomes e nomes principais, como jose, silva, etc*

"**qtd_sigla_uf**": *0*, *# quantidade de siglas de unidades da federação*

"**qtd_nome_uf**": *0 # quantidade de nomes de unidades da federação*

},

"**nome_arquivo_filtrado**": "*carregando modelos*", *# nome do arquivo filtrado, após pré-processamento, somente com palavras válidas, sem acento e em minúscula*

"**texto_filtrado**": "*execução inicial apenas*" *# texto filtrado, após pré-processamento, somente com palavras consideradas válidas e nomes de classes substitutas (cpf, data, etc), sem acento e em minúscula*

}

4.4 RASTRO NO CLADOP

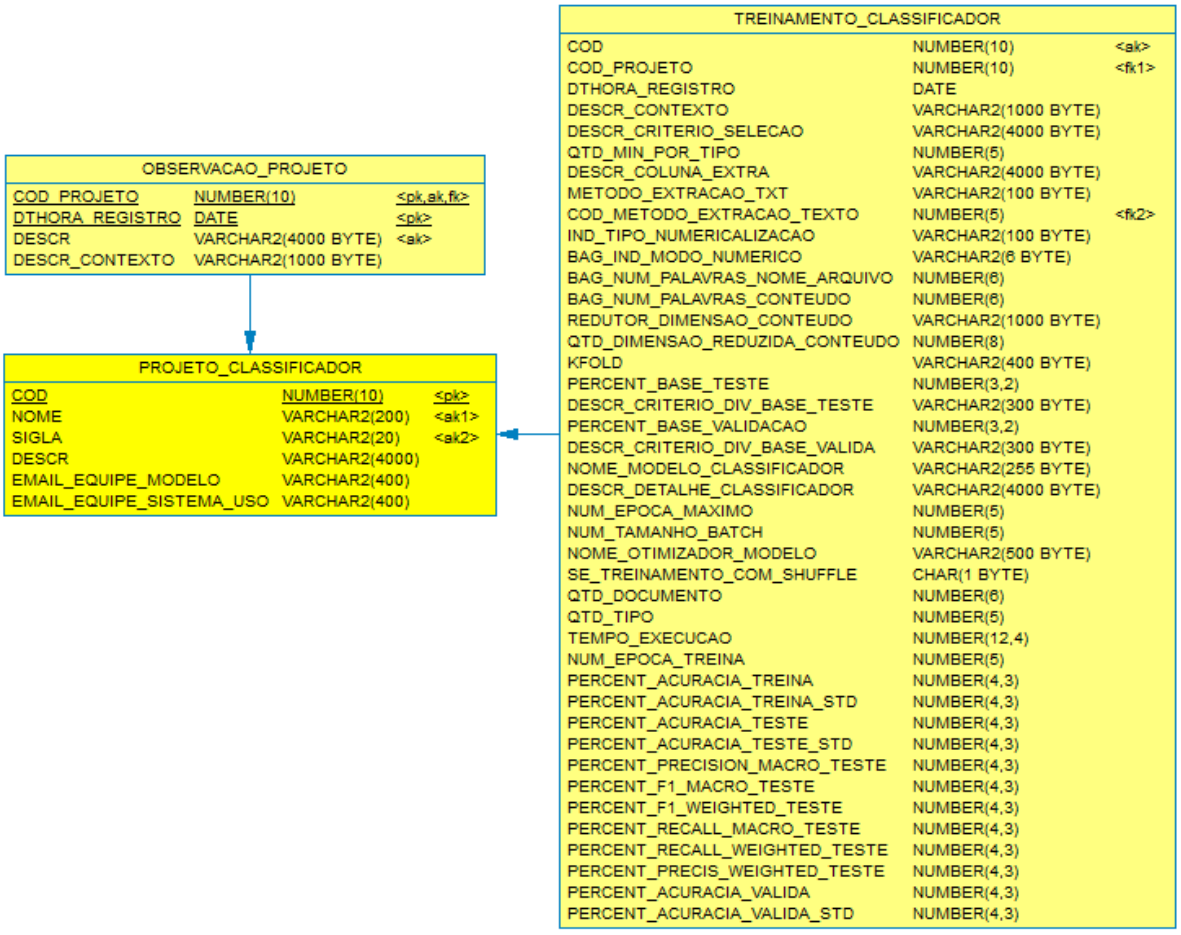
Seguem detalhes de como as atividades do Rastro-DM foram implementadas no projeto[30].

30 Os dados do rastro no Cladop e o código usado para sua construção encontram-se publicados em https://gitlab.com/MarcusBorela/rastro-dm.git, na pasta Rastro_Projeto_Cladop.

4.4.1 Definição de Ação

As Definições de Ação foram registradas em uma tabela[31] de banco de dados denominada *observacao_projeto* constante do Modelo Entidade Relacionamento da Figura 8 que apresenta as tabelas que persistem o rastro do projeto Cladop.

Figura 8 - Modelo Entidade Relacionamento com tabelas que persistem dados do rastro no Cladop



Fonte: elaborada pelo autor (2019)

A Tabela 10 ilustra algumas definições de ação registradas durante o projeto, com o momento do registro, o contexto[32] e a tarefa CRISP-DM associada, que foi indicada

[31] As primeiras definições de ação registradas, de menor complexidade, foram no código pois a sua execução se dava no próprio caderno, no espaço que se abria à frente da atividade. Porém, com o tempo, elas passaram a ser mais complexas e amplas e não mais implementada no código, mas em vários códigos extras. Definitivamente não se pode confiar o registro de definições de ação a código. Devido à natureza imprevisível de um projeto DM o código muitas vezes evolui rapidamente e acaba comprometendo, entre outras coisas, sua documentação (GREFF et al, 2017).

[32] Indica o caderno (*ipython notebook*) em que se executava o comando *registrar_historico* (<texto>).

posteriormente ao registro e ilustra a iteração das atividades em um processo de DM discutida no Referencial Teórico: uma tarefa é executada diversas vezes no curso do projeto.

Tabela 6 – Exemplos de definições de ação registradas no Cladop que tratam apenas de fluxo de projeto

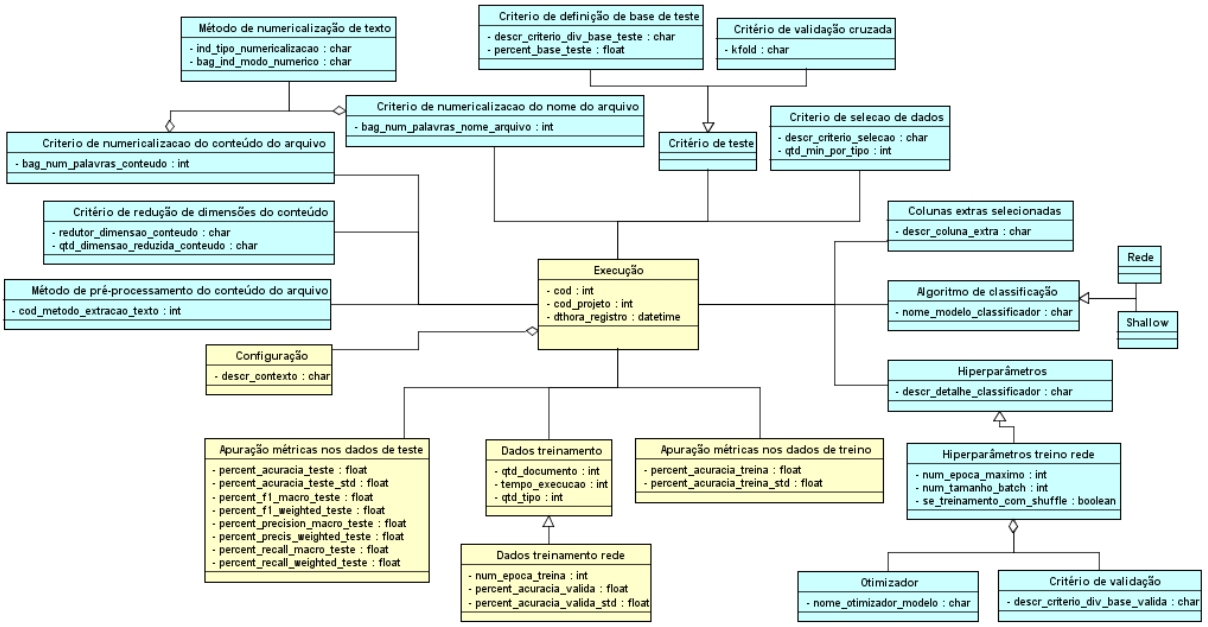
Momento	Descrição	Contexto	Tarefa CRISP-DM
12/3/2019 17:04	Criando estrutura (código e dados) para tratar *k-fold* em *shallow algorithms*	Cladop-colunas_binárias.ipynb	Projeto de testes
14/3/2019 19:27	Iniciei execuções para experimentar otimizadores (MLP): *nadam*, *adadelta*	Cladop.ipynb	Construir Modelo
18/3/2019 09:01	Vou testar usando valores normais nas colunas discretas (*cod_unidade*, etc) ao invés de várias colunas binárias. Será que melhora para *shallow*?	Cladop-Copy1.ipynb	Formatar Dados
18/3/2019 11:40	Experimentando colunas não mais binárias com MLP	Cladop-colunas_binárias.ipynb	Formatar Dados
26/3/2019 11:57	Iniciando inclusão de nomes de arquivos no modelo	Cladop.ipynb	Selecionar Dados
3/4/2019 17:51	Experimentando ensemble com AdaBoost sobre LGBMClassifier	Cladop.ipynb	Construir Modelo
10/4/2019 11:27	Passamos a gravar no log dados de teste das métricas: "*precision*_macro", "*precision_weighted*", "f1_macro", "f1_*weighted*", "recall_macro", "recall_*weighted*". Mas, em um primeiro momento, somente para algoritmos *shallow* com *kfold*.	Cladop-Copy2.ipynb	Projeto de testes
24/4/2019 19:53	Experimentando variações com o criterio de colunas extras para lgbm	Cladop-Copy1.ipynb	Selecionar Dados
29/5/2019 19:30	Alterado programa (*shallow*) para gravar também *recall* e f1 micro	Cladop-Copy1.ipynb	Projeto de testes
29/5/2019 20:46	Experimentando *lightgbm* após ajuste nos tipos - *se_ativo*=n e excluído	Cladop-Copy2.ipynb	Construir Modelo
2/6/2019 17:26	Usando critério de divisão de base de validação em caso de *kfold* (*rede MLP*), ao invés de percentual fixo.	Cladop-rede-nome_arquivo-Copy1.ipynb	Construir Modelo

Fonte: elaborada pelo autor (2019)

4.4.2 Registro de Treinamento

A Figura 9 apresenta os dados que são persistidos em cada treinamento agrupados pelos principais conceitos envolvidos. Em azul claro, constam os dados que são entrada para o treinamento e em amarelo os dados derivados do processo[33].

Figura 9 - Informações persistidas de um treinamento no projeto Cladop agrupadas por conceito



Fonte: elaborada pelo autor (2019)

Todos os dados do treinamento são persistidos em uma única tabela *treinamento_classificador*, conforme o Modelo Entidade Relacionamento da Figura 8.

A Tabela 7 traz a descrição dos dados de entrada, o conceito associado da Figura 8 e a tarefa CRISP-DM que lhes dão origem.

Tabela 7- Dados do rastro de entrada para um treinamento no Cladop

Tarefa CRISP-DM	Conceito	Parâmetro	Descrição
Construir modelo	Hiperparâmetros	*descr_detalhe_classificador*	Descrição dos hiperparâmetros da execução do treinamento. Ex.: *n_estimators*, *learning_rate*.
Construir modelo	Hiperparâmetros Treino Rede	*num_epoca_maximo*	Existe se o classificador for *rede MLP*. Número máximo de épocas passadas como parâmetro para o treinamento.
Construir modelo	Hiperparâmetros Treino Rede	*num_tamanho_batch*	Existe se o classificador for *rede MLP*. Tamanho do batch usado para treinamento.

[33] Não foi colocado no diagrama UML a cardinalidade dos relacionamentos por simplificação. Pode-se assumir que um treinamento tem no máximo um dos atributos. A ausência de alguns atributos é justificada pela evolução das técnicas experimentadas durante o projeto.

Construir modelo	Hiperparâmetros treino rede	*se_treinamento_com_shuffle*	Existe se o classificador for *rede MLP*. Indica se os dados devem ser misturados antes de cada nova época de treinamento.
Construir modelo	Otimizador	*nome_otimizador_modelo*	Existe se o classificador for *rede MLP*. Otimizador usado. Armazenado como um objeto de classe em *python*, por exemplo: *keras.optimizers.Adam* e os seus parâmetros.
Construir modelo	Critério de validação	*percent_base_validacao*	Existe se o modelo for *rede MLP*. Percentual dos dados reservado para base de validação no treinamento da rede.
Gerar Novos Dados	Critério de redução de dimensões do conteúdo	*qtd_dimensao_reduzida_conteudo*	Indica a quantidade de dimensões que o *bag do conteúdo* terá após redução de dimensionalidade. Opcional.
Gerar Novos Dados	Critério de redução de dimensões do conteúdo	*redutor_dimensao_conteudo*	Indica o redutor a ser usado para reduzir dimensionalidade do *bag* do conteúdo do arquivo. Armazenado como objeto da classe em *python*, por exemplo, TruncatedSVD(), e seus parâmetros. Opcional.
Formatar Dados	Método de substituição de texto por números	*ind_tipo_numericalizacao*	Indica o tipo de substituição de texto por números aplicado tanto ao conteúdo quanto ao nome do arquivo. Na versão atual, utiliza-se *bag of words.*[34]
Formatar Dados	Método de substituição de texto por números	*bag_ind_modo_numerico*	Existe se o *ind_tipo_numericalizacao* for *bag of words*, indica o método para definição do número associado a cada palavra na geração da matriz de um documento. Pode ser: *binary, count, tfidf* ou *freq*. Aplicado tanto ao conteúdo quanto ao nome do arquivo.
Formatar Dados	Criterio de numericalizacao do nome do arquivo	*bag_num_palavras_nome_arquivo*	Indica número de palavras usado para geração da matriz de um documento a partir do nome do arquivo.
Formatar Dados	Criterio de numericalizacao do conteúdo do arquivo	*bag_num_palavras_conteudo*	Existe se o *ind_tipo_numericalizacao* for *bag of words*, indica o número de palavras usado para geração da matriz de um documento a partir do conteúdo do arquivo.
Projeto de testes	Critério de validação cruzada	*Kfold*	Indica a forma de validação cruzada para apuração da acurácia (micro). Armazenado como um objeto de classe em *python, por exemplo, ShuffleSplit() e kfold()* e os seus parâmetros.
Projeto de testes	Criterio de definição de base de teste	*percent_base_teste*	Existe se não for informado *kfold*. Percentual dos dados reservado para base de teste antes do treinamento.
Projeto de testes	Criterio de definição de base de teste	*descr_criterio_div_base_teste*	Existe se não for informado *kfold*. Indica critérios adicionais usados para construção da base de teste. Ex.: com/sem *shuffle*; com/sem estratificação.
Selecionar Dados	Colunas extras selecionadas	*descr_coluna_extra*	Indica colunas extras a serem consideradas pelo modelo, como características da qualidade do arquivo ou variáveis de contexto do Dano associado ao documento. Armazenado em formato sql (*Structured Query Language*) para ser concatenado ao comando de busca dos dados.
Selecionar Dados	Criterio de selecao de dados	*qtd_min_por_tipo*	Quantidade mínima exigida por tipo para treinamento.

[34] Ficou fora do escopo do projeto experimentar MLP com representação numérica das palavras por *embeddings,* pré-treinados, ao invés de *bag of words*. Parece uma ideia promissora pois os *embeddings* levam em consideração o significado dos termos.

Selecionar Dados	Criterio de selecao de dados	*texto_criterio_seleção*	Critério usado para seleção (filtro) dos dados. Armazenado em formato *sql* para ser concatenado ao comando de busca de dados.
Selecionar Dados	Método de pré-processamento do conteúdo do arquivo	*cod_metodo_extracao_texto*	Método usado para pré-processamento do texto, ou seja, para a extração de palavras válidas do documento PDF. [35]
Selecionar técnicas	Algoritmo	*nome_modelo_classificador*	Nome do modelo de classificador usado. Foram experimentados[36] no contexto do projeto algoritmos *shallow* (como *RandonForest* e *lightGBM*) e redes MLP (*Multi Layer Perceptron* da biblioteca *Keras.Sequential*).

Fonte: elaborada pelo autor (2019)

No rastro do Cladop, por falta de necessidade, não foram persistidos os dados usados nas experimentações, mas apenas os parâmetros usados para a seleção das variáveis e para filtro dos dados. Contudo, os parâmetros de seleção de dados usados e a configuração da base espelho com os dados para treinamento[37] permitem sua geração a qualquer momento.

Conforme discutido na Seção 3.3, além dos parâmetros de treinamentos e de testes e dos dados usados, que são entrada para um treinamento, são envolvidos outros 3 grupos de informações que são saída do treino: resultados obtidos, dados de configuração e dados de contexto. A Tabela 8 descreve os dados registrados automaticamente em cada treinamento no projeto Cladop, separando-os por grupo e conceito da Figura 8.

Tabela 8 - Dados do rastro gerados em um treinamento no Cladop

Grupo	Conceito	Item	Descrição
Dados de Configuração	Configuração	*descr_contexto*	Indica o contexto de execução, no caso o nome do programa *python* ou do caderno *ipython*. Permite, em conjunto com a data da execução, identificar a versão do código executado na ferramenta de controle de versões de código.
Dados de Contexto	Dados treinamento rede	*num_epoca_treina*	Existe se o algoritmo for *rede MLP*. Número de épocas efetivas de treinamento para se chegar ao modelo com melhor acurácia.
Dados de Contexto	Execução	*cod*	Identifica de forma única uma execução de treinamento do modelo, e, dessa forma, um modelo gerado pela execução. Chave sequencial do treinamento na tabela no banco de dados.
Dados de Contexto	Execução	*cod_projeto*	Identifica o projeto associado ao treino.

[35] Foram experimentados sete métodos, combinações diferentes de 3 processos: obtenção do texto (OCR original ou se executado novo OCR via *Tesseract*), critério de validação se um termo é uma palavra válida e classes de palavras consideradas para substituição de termos.

[36] Outras técnicas mais modernas podem ser experimentadas no futuro. Contudo, como a sequência das palavras não traduz a realidade do texto (OCR), não se espera que técnicas mais modernas, que usam redes recorrentes, que se baseiam no sequenciamento, alcancem acurácia muito superior.

[37] A existência de algumas datas nessa base (de replicação e de pré-processamento dos textos) permite a composição de uma lógica de geração de uma base usada em um treinamento.

Dados de Contexto	Execução	*dthora_registro*	Momento de registro do rastro no banco de dados ao final da execução.
Dados de Contexto	Dados treinamento	*qtd_documento*	Quantidade de documentos usados no treinamento, que satisfizeram os critérios de seleção de dados.
Dados de Contexto	Dados treinamento	*qtd_tipo*	Quantidade de tipos considerados no treinamento, que satisfizeram os critérios de seleção de dados.
Dados de Contexto	Dados treinamento	*tempo_execucao*	Indica o tempo de execução em segundos da ação de fit de treinamento do modelo[38].
Resultados obtidos	Apuração métricas nos dados de teste	*percent_acuracia_teste*	Percentual de acurácia alcançado na base de teste.
Resultados obtidos	Apuração métricas nos dados de teste	*percent_acuracia_teste_std*	Existe se houver validação cruzada (*kfold*). Desvio padrão da acurácia do classificador nos dados de teste
Resultados obtidos	Apuração métricas nos dados de teste	*percent_f1_macro_teste*	Percentual da métrica f1_macro apurado na base de teste. [39]
Resultados obtidos	Apuração métricas nos dados de teste	*percent_f1_micro_teste*	Percentual da métrica f1_micro apurado na base de teste.
Resultados obtidos	Apuração métricas nos dados de teste	*percent_f1_weighted_teste*	Percentual da métrica f1_*weighted* apurado na base de teste.
Resultados obtidos	Apuração métricas nos dados de teste	*percent_precis_weighted_teste*	Percentual da métrica *precision_weighted* apurado na base de teste.
Resultados obtidos	Apuração métricas nos dados de teste	*percent_precision_macro_teste*	Percentual da métrica precison macro apurado na base de teste.
Resultados obtidos	Apuração métricas nos dados de teste	*percent_recall_macro_teste*	Percentual da métrica *recall*_macro apurado na base de teste.
Resultados obtidos	Apuração métricas nos dados de teste	*percent_recall_micro_teste*	Percentual da métrica *recall*_micro apurado na base de teste.
Resultados obtidos	Apuração métricas nos dados de teste	*percent_recall_weighted_teste*	Percentual da métrica *recall*_weighted apurado na base de teste.
Resultados obtidos	Apuração métricas nos dados de treino	*percent_acuracia_treina*	Percentual de acurácia alcançado na base de treinamento.
Resultados obtidos	Apuração métricas nos dados de treino	*percent_acuracia_treina_std*	Existe se houver validação cruzada (*kfold*). Desvio padrão da acurácia do classificador nos dados de treinamento.
Resultados obtidos	Dados treinamento rede	*percent_acuracia_valida*	Existe se o classificador for *rede MLP*. Percentual de acurácia na base de validação.
Resultados obtidos	Dados treinamento rede	*percent_acuracia_valida_std*	Existe se o classificador for *rede MLP* e se houver validação cruzada (*kfold*). Desvio padrão da acurácia do classificador nos dados de validação.

Fonte: elaborada pelo autor (2019)

Cabe comentar a ausência de dois importantes itens no rastro: o modelo propriamente dito e a mensagem de erro em caso interrupção de execução. Não tem mensagem de erro, pois apenas foram registrados os treinamentos com sucesso[40]. Quanto ao modelo, somente houve a necessidade de se gerar o classificador para as versões implantadas, logo as versões

[38] No caso de validação cruzada, representa o tempo de execução do *fit* do último modelo.

[39] As métricas f1, recall e precisão são calculados só para algoritmos *shallow* pois serem retornadas no método *cross_validate* do *scikit-learn* e não ter sido programado esse cálculo para o caso de redes usando *Keras*.

[40]Além de permitir o aprendizado sobre erros e ações no tratamento dos erros mais frequentes, essa informação também pode ser um indicativo da qualidade do código desenvolvido para o treino.

intermediárias não foram salvas, o que, com certeza, reflete um rastro menor em número de bytes usados.

Uma versão de um Classificador está associada a dois treinamentos: um para apuração das métricas, de preferência usando validação cruzada para separar os dados de teste, e outro para a construção efetiva do modelo, em que todos os dados podem ser usados para o treinamento, sem a necessidade de se separar dados para testes, o que tende a gerar um modelo melhor. As chaves dos rastros (coluna *cod* da tabela *treinamento_classificador*) foram usadas para indicar as versões do classificador nos metadados do projeto, em uma tabela de banco de dados denominada *versao_classificador*, conforme Tabela 9.

Tabela 9 - Metadados das versões do Classificador com referência ao rastro

Coluna	Registro 1	Registro 2	Registro 3
num_versao	1	2	2
num_subversao	0	0	1
cod_treinamento_teste		13.137	13.134
cod_treinamento_geracao	1.074	13.130	13.138
dthora_implantacao	01/21/2019 10:01:36	07/01/2019 00:00:00	

Fonte: elaborada pelo autor (2019)

A Tabela 10 ilustra o rastro dos treinamentos de teste e de geração associadas à versão 2.1 do Cladop, cujos códigos são, respectivamente, 13.134 e 13.138. Alguns itens não constam da relação pois não têm valores, uma vez que foi realizado teste com validação cruzada (justifica ausência dos itens *percent_base_teste* e *descr_criterio_div_base_teste*), não foram usadas colunas adicionais ao nome e ao conteúdo do arquivo (*descr_coluna_extra*) e o algoritmo usado é de redes neurais (percentuais de precisão, *recall* e f1, na forma macro e *weighted*).

Tabela 10 - Rastro dos treinamentos de teste e de geração da versão 2.1 do Cladop

Grupo	Item	Teste	Geração
Contexto	*Cod*	13.134	13.138
	cod_projeto	1	
	dthora_registro	05/07/2019 00:44:59	05/07/2019 18:02:12
	num_epoca_treina	33	24
	qtd_documento	63.468	
	qtd_tipo	81	
	tempo_execucao	460	475
Configuração	*descr_contexto*	Cladop_monitoramento.ipynb	
Parâmetros	*bag_ind_modo_numerico*	Tfidf	
	bag_num_palavras_conteudo	24.576	
	bag_num_palavras_nome_arquivo	1.000	
	cod_metodo_extracao_texto	7	
	descr_criterio_selecao	TRUNC (doc.data_criacao) > to_date('1/5/2018', 'dd/mm/yyyy') and doc.cod_tipo <> 41 -- outros	

	descr_detalhe_classificador	7- validacao estratificada, rede 2 camadas (1024_relu_dropout50 e 1024_sigmoid_dropout50) reduceLR50[41]	
	ind_tipo_numericalizacao	Bag of words	
	Kfold	Repeated*Kfold*(n_splits=7, n_repeats=2, random_state=42)	(não separados dados para teste)
	nome_modelo_classificador	Rede - MLP	
	nome_otimizador_modelo	<class 'keras.optimizers.Adam'> Parâmetros: {'lr': 9.999999747378752e-05, 'beta_1': 0.8999999761581421, 'beta_2': 0.9990000128746033, 'decay': 0.0, 'epsilon': 1e-07, 'amsgrad': False}	
	num_epoca_maximo	100	
	num_tamanho_batch	256	
	qtd_dimensao_reduzida_conteudo	768	
	qtd_min_por_tipo	0	
	redutor_dimensao_conteudo	TruncatedSVD(algorithm="arpack", n_components=768, n_iter=5, random_state=42, tol=0.0)	
	se_treinamento_com_shuffle	N	
Resultado	*descr_criterio_div_base_valida*	sem estratificacao com *shuffle*	
	percent_acuracia_teste	91,1%	
	percent_acuracia_teste_std	0,3%	
	percent_acuracia_treina	96,4%	95,6%
	percent_acuracia_treina_std	0,7%	
	percent_acuracia_valida	90,8%	92,1%
	percent_acuracia_valida_std	0,6%	
	percent_base_validacao	5,0%	5,0%

Fonte: elaborada pelo autor (2019)

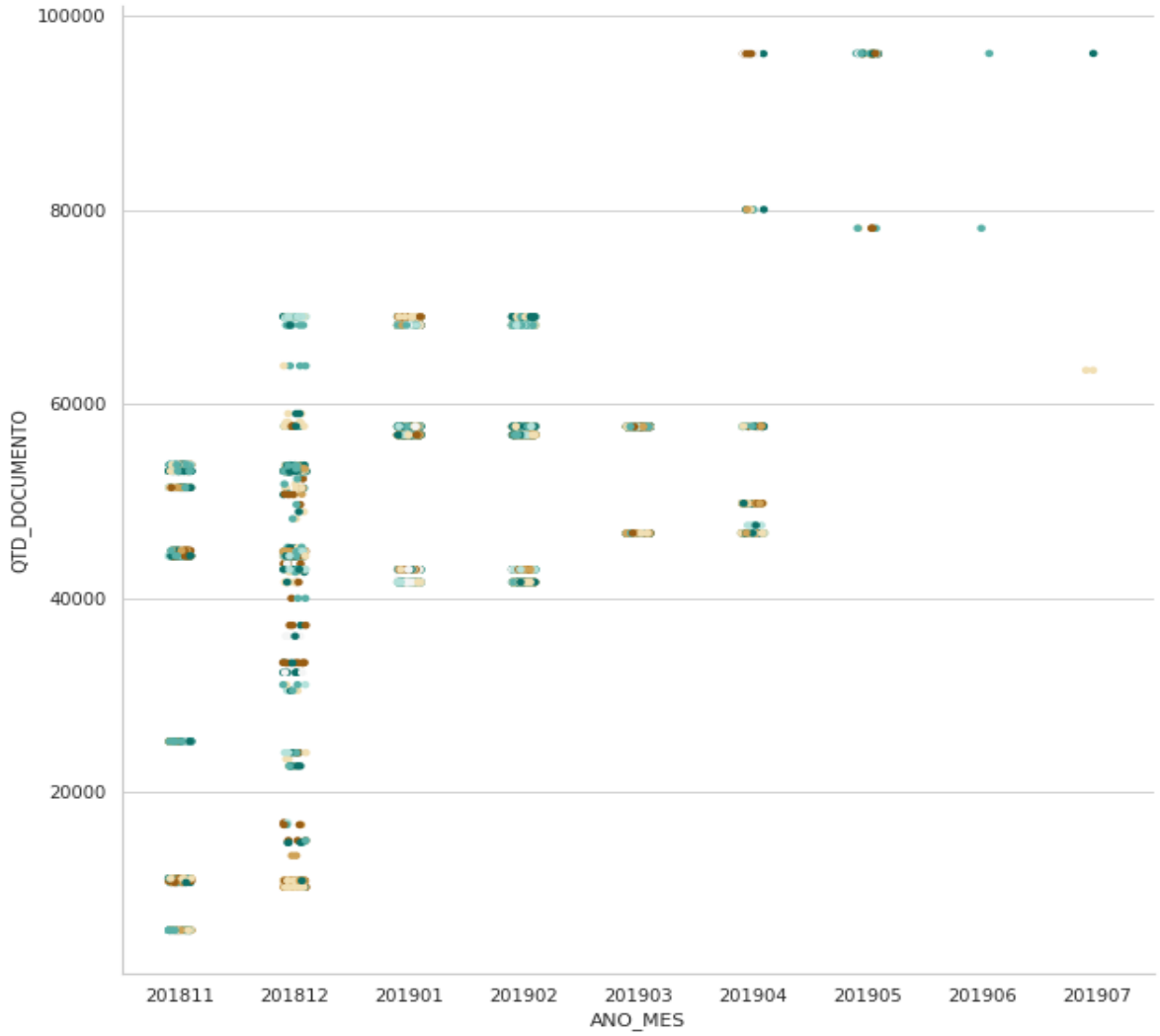
O Gráfico 4 mostra a quantidade de treinamentos mês a mês durante o projeto (eixo x) e a quantidade de documentos usados nas experimentações. A cor do ponto indica a acurácia micro alcançada na base de testes: quanto mais verde, maior, quanto mais marrom, menor. Percebe-se que a quantidade de documentos aumentou com o passar do tempo, pois novos documentos foram inseridos no sistema e-TCE[42]. Alguns treinamentos envolvendo documentos com alta qualidade de OCR e tipos com grande quantidade de exemplares levaram a classificadores com acurácias de 98% em base de teste[43]. Por terem critérios mais restritivos, envolveram menos documentos. No decorrer do projeto, optou-se por construir um classificador que abrangesse todos os tipos de documentos, independentemente da quantidade de documentos por tipo.

[41] Rede gerada pelos commandos:
model.add(Dense(1024, input_shape=(qtd_colunas,), activation='relu'));
model.add(Dropout(0.5)); model.add(Dense(1024, activation='sigmoid')); model.add(Dropout(0.5))
model.add(Dense(treinamodelo['qtd_tipo'], activation='softmax')); model.compile(loss='categorical_crossentropy', metrics=["accuracy"])

[42] Foram realizadas três cargas na base espelho de treinamento com as seguintes quantidades de documentos: 53.914 (12/11/2018), 69.330 (11/12/2018) e 108.292 (18/4/2019).

[43] O treino de código 147 em 24/11/2018 alcançou 98% envolvendo 11110 documentos selecionados com o critério *(qtd_token_palavra_valida/ava.qtd_pagina_pdf) >= 30 and doc.cod_tipo in (42,43,44,45)*

Gráfico 4 – Evolução mês a mês de treinamentos, com quantidade de documentos e a acurácia na base de teste[44].



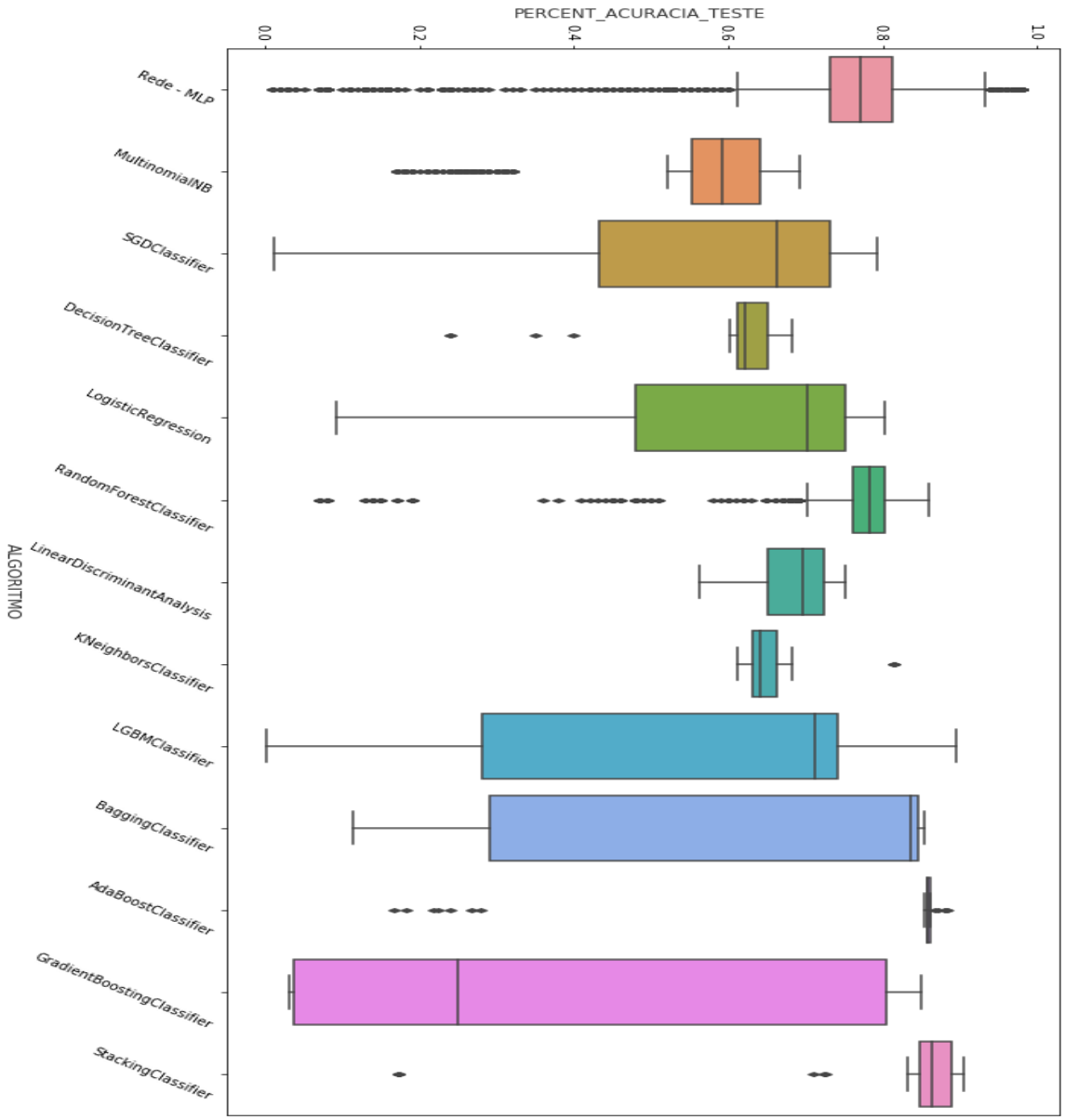
Fonte: elaborada pelo autor (2019)

Durante o projeto foram experimentados vários algoritmos. Os modelos com maior acurácia implementaram redes neurais. Modelos implementados usando o algoritmo *lightGBM* ficaram logo atrás, com uma diferença de cerca de 2%, mantidos iguais todos os demais parâmetros. A arquitetura de rede usada está descrita na tabela anterior. O Gráfico[45] 5 mostra a acurácia em base de teste alcançada pelos algoritmos experimentados[46].

[44] Comando *python* usado para sua geração: seaborn.catplot(data=df, x="ANO_MES", y="QTD_DOCUMENTO", height=10, aspect=1, row_order=True, hue='PERCENT_ACURACIA_TESTE', legend=False, , palette=sns.color_palette("BrBG", 7))

[45] Comando usado: *seaborn.boxplot(x="NOME_MODELO_CLASSIFICADOR", y="PERCENT_ACURACIA_TESTE", data=df, ax=ax)*

Gráfico 5 - *Boxplot* com acurácia em base de teste dos algoritmos experimentados.



Fonte: elaborada pelo autor (2019)

[46] A grande variação de acurácia por algoritmo se justifica pela variação de combinação de parâmetros usados nos treinamentos, como o critério de seleção de documentos envolvidos, as diversas combinações de valores de hiperparâmetros das próprias técnicas e os critérios de teste.

4.4.3 Síntese de Aprendizado

Os aprendizados foram armazenados também como observações na tabela *observacao_projeto*. A Tabela 11 ilustra alguns aprendizados importantes do projeto. Algumas observações mesclam definições de ação e aprendizados, sendo os aprendizados resultados das ações ou motivadores delas[47].

Tabela 11 – Exemplos de aprendizados registrados durante o projeto

Momento	Descrição	Atividade CRISP-DM
27/2/19 10:32	Somente as variáveis extras selecionadas (de qualidade e de contexto) levam a um resultado de 44% de acurácia, com otimizador Adam (se Adagrad:42%, se SGD:28%)	*Selecionar Dados*
27/2/19 10:39	Ao usar *pca* (*sklearn*): melhor separar o comando *fit* do comando *transform*. O comando *fit_transform* estava travando!	*Formatar Dados*
27/2/19 10:53	Ao usar *pca* (*sklearn*): se o número de dimensões for bem pequeno e o *array* for grande (ver detalhes na documentação da função), melhor usar o método *randomized* do que o *full*. Pois é mais rápido, e os resultados (variância alcançada) são equivalentes.	*Formatar Dados*
1/3/19 18:14	Considerando *randomforest* (sem tipo outros): O bag menor (86) sem variáveis extras já traz um resultado interessante: 71%; com variáveis extras, há um aumento para 75,10%. MLP aumenta para 75,85%; as 10 palavras mais comuns sem variáveis extras levam a acurácia de 50.80%. Com variáveis extras: 66.5%.	*Selecionar Dados*
1/3/19 18:22	*Truncated SVD* com tamanho 43 (metade do número de classes) se mostrou mais benéfico (execução rápida e melhor acurácia - com *randomforest*) do que *SparcePCA* (muito lento!) e *RBM* (tem baixa acurácia)	*Formatar Dados*
14/3/19 13:11	Precisarei apagar as execuções com *k-fold*, pois a acurácia baixa provavelmente se dava por causa do algoritmo receber y com várias colunas (pensaria ser problema *multilabel* e não multiclasse), quando esperam receber apenas uma coluna.	*Projeto de testes*
15/3/19 15:24	Experimentando teste com *repeated kfold*: como o treinamento recebe como parâmetro o objeto "*kfold*", bastou mudar a classe.	*Projeto de testes*
15/3/19 21:05	Melhor usar *sklearn.model_selection.cross_validate* do que *cross_val_score* pois ele traz também outras informações como métrica em treinamento (se solicitado). Alterado código para gravar treinamento com teste usando *kfold* para algoritmos *shallow*.	*Projeto de testes*
19/3/19 9:29	Tratado *encoding* dos dados vindos do *oracle* (*cladop.utilbd.connecta_bd* com UTF-8)	*Formatar Dados*
29/03/2019 09:54	Percebida uma melhora de cerca de 5% na acurácia dos modelos após inclusão do nome do arquivo como característica.	*Selecionar Dados*
1/4/19 15:36	Para algoritmos *shallow*, as colunas extras com valores não *dummies* (uma só coluna com vários valores discretos) levaram a um resultado melhor. Para Rede Neural, há uma pequena melhora usando valores *dummies*..	*Formatar Dados*
3/4/19 17:29	*Lightdbm.sklearn.LGBMClassifier* com *lerning rate* 0.5 ou maior não alcançou acurácia superior a 40%.	*Construir Modelo*
3/4/19 17:47	Aplicando *boost* (*adaboost*) sobre o *RandonForest* alcançou-se uma melhoria na acurácia de aproximadamente 1% - verificação visual	*Construir Modelo*
10/4/19 10:49	Em classificação de multiclasses, as métricas *f1_micro, recall_micro, precision_micro* e acurácia são equivalentes	*Projeto de testes*

[47] Exemplo de aprendizado que motiva uma definição de ação: Alterado programa (*shallow*) para não gravar *recall* e f1 micro, pois são equivalentes a acurácia (e precisão), registrado em 6/4/2019.

10/4/19 10:53	Em classificação de múltiplas classes, as métricas f1, *recall* e *precision*, quando calculadas com peso (*weighted*) chegam a valor bem próximo de cálculo por instância (calculo "micro" - que equivale a acurácia)	*Projeto de testes*
12/4/19 9:56	Criado modelo 7 de tratamento de texto com classes novas: nome, sobrenome, sigla uf e nome uf. Bem como integrado com as palavras mais comuns no modelo de nomes de arquivos. Também feita tradução de palavras comuns com erro: aenxo por anexo	*Formatar Dados*
15/5/19 7:44	Para a rede 7, o *binary* demorou mais (36 contra 33 épocas) e teve acurácia um pouco pior. *Tfidf*: 89,9 +- 0,3 contra 89,5 +- 0.2.	*Formatar Dados*
16/5/19 17:34	Otimizador em redes neurais passou a ser um objeto, para se ter flexibilidade de ajustar parâmetros	*Construir Modelo*
22/5/19 20:04	O otimizador *Adam* (passando objeto com *Amsgrad=False*) foi o melhor otimizador ate o momento. Acima do A*msgrad=True* em 0.1% , do *Rmsprop* e do *Adadelta* em 0.2% no contexto avaliado das últimas execuções.	*Construir Modelo*
29/5/19 20:43	Em 11/04/2019 todos os 830 tipos 86 (*Notificação, inclusive edital - responsabilidade afastada*) viraram 42 [*Notificação (ofício), inclusive edital*]. Atualizamos tabela espelho para refletir essa mudança.	*Coletar Dados*
5/6/19 18:20	Passamos a considerar para nome de arquivo um bag de 1000 palavras, pois constatamos pelo *lgbmClassifier.booster_.feature_importance()* que a partir de 1000 a importância era zero. E *rede MLP* melhorou 0.15% sua acurácia.	*Formatar Dados*
5/6/19 19:57	Usando modelo de rede 7 com *shuffle==True* não impactou a acurácia. Na realidade reduziu em 0.04%.	*Construir Modelo*

Fonte: elaborada pelo autor (2019)

Há aprendizados da tabela, a maioria, que podem muito bem ser sintetizados por mineração de dados nos registros de treinamentos. Por exemplo, o primeiro aprendizado do dia 10/4/2019: facilmente esse padrão, igualdade de valores em algumas colunas, poderia ser detectado por uma rotina de pesquisa de regras nos dados. Mas há alguns que carecem da intervenção humana, como o primeiro do dia 15/3/2019: *Experimentando teste com repeated kfold: como o treinamento recebe como parâmetro o objeto 'kfold', bastou mudar a classe*.

Não foi escopo deste trabalho e nem do projeto Cladop a geração de conhecimentos de forma automática. Mas, esse se mostra um caminho promissor para a elevação do conhecimento organizacional, o que foi referendado por NGUYEN (2018), que, como visto, ressalta haver um grande vazio de pesquisas nessa área.

Passados menos de seis meses desde que os primeiros registros constantes da Tabela 11 foram efetuados, o autor experimentou contabilizar quantos deles ainda se encontram completamente claros em sua memória e chegou ao número de 7[48]entre os 22, um total de 27%. Os demais, infelizmente, estavam perdidos da memória do autor, mas, graças ao rastro, não perdidos da memória do projeto.

Esse esquecimento leva à execução repetida de experimentos (BECKER e GHEDINI, 2005). Durante o projeto Cladop houve um esquecimento pela equipe que levou a uma execução

[48] São eles os registros dos dias 15 (15h) e 29 de março, 1, 10 (10:49) e 12 de abril e 16 de maio.

repetida de tarefa. Aconteceu uma vez no projeto[49]: o aprendizado *Em classificação de multiplas classes, as métricas f1_micro, recall_micro, precision_micro e acurácia são equivalentes*, de 10/4/2019, foi ignorado, no dia 29/5/2019, com a execução da ação definida como: *Alterado programa (shallow) para gravar também recall e f1 micro*, o que levou a um retrabalho em 6/4/2019: *Alterado programa (shallow) para não gravar recall e f1 'micro', pois são equivalentes a acurácia (e precisão).* Ficou a lição aprendida de que não basta ter o rastro, é necessário fazer uso do mesmo.

4.5 GERAÇÃO DE RELATÓRIO A PARTIR DE INFORMAÇÕES NO RASTRO

Um importante uso para o rastro percebido no projeto é a possibilidade de construção de forma semi-automática de relatórios que podem ser previstos na *metodologia base* em uso na Organização. Ainda que sem todos os detalhes necessários, uma versão preliminar do relatório pode ser gerada de forma automática a partir do rastro.

O relatório criado como exemplo é sobre a substituição de palavras por números (*numericalization*) dos textos no projeto Cladop, passo que integra a fase *Preparação de dados* do CRISP-DM e perpassa as atividades de *Seleção*, *Construção* e *Formatação de dados.*

A partir de gráficos gerados automaticamente dos treinamentos do rastro do Cladop envolvendo parâmetros relacionados à substituição de palavras por números, foram adicionadas informações que os contextualizam. Em seguida, foram relacionados os registros de aprendizados e de definições de ação em ordem cronológica sobre o tema. Faz parte do relatório uma introdução que está associada ao tema do relatório e pode ser gerada também de forma automática.

A lista cronológica de aprendizados e definições de ação e alguns gráficos gerados automaticamente a partir dos treinamentos formam um bom ponto de partida para um relatório sobre a alguma tarefa específica da *metodologia base* que, normalmente, pretende detalhar decisões e suas razões, suposições, descobertas e conhecimentos adquiridos na execução. Segundo BECKER e GHEDINI (2005), o CRISP-DM solicita essas informações, embora não forneça uma estrutura para relatório.

[49] Espera-se que outras não tenham caído no esquecimento!

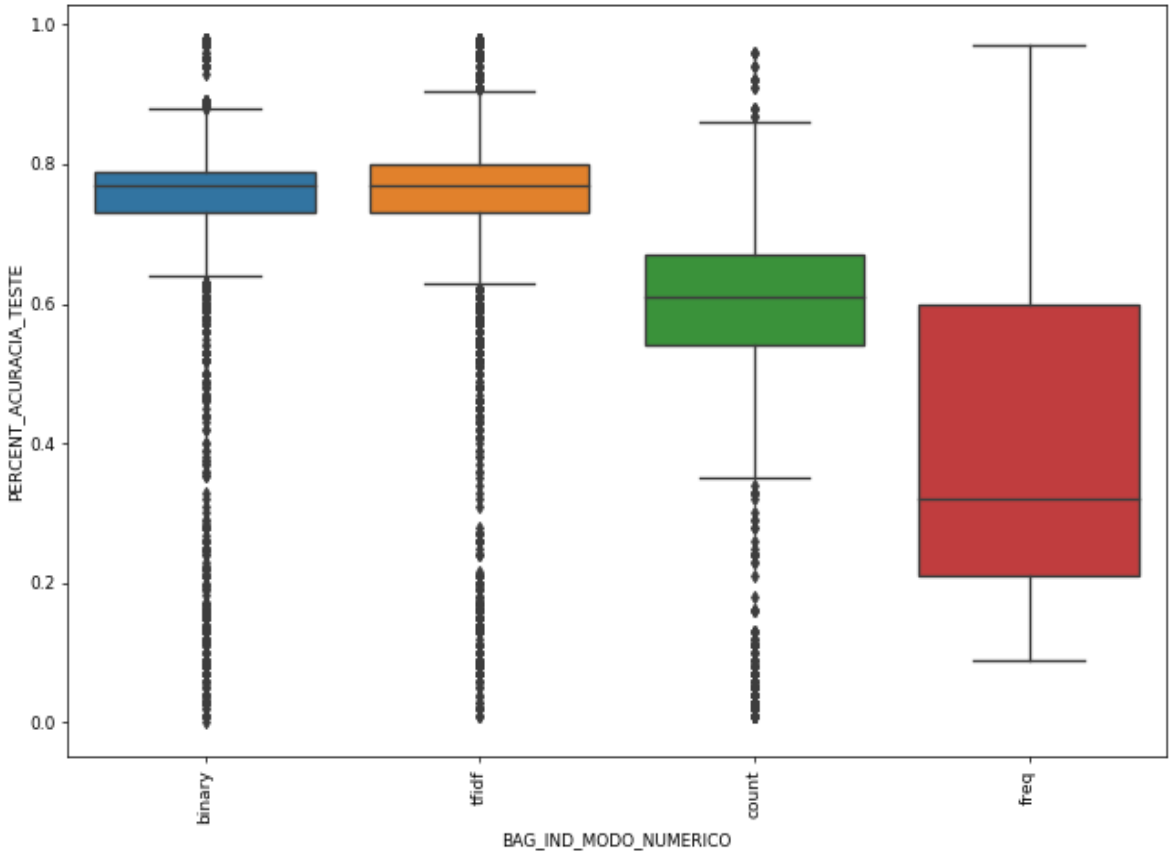
Relatório sobre a substituição de palavras por números no Cladop

Essa atividade integra a fase *Preparação de dados* do CRISP-DM e perpassa as atividades de *Seleção*, *Construção* e *Formatação de dados.*

Os textos precisam ser transformados em números para que possam ser usados em algoritmos de máquina. No projeto foi escolhida a técnica *bag of words* que, resumidamente, associa um número a cada palavra do *corpus*, conjunto total de textos tratados, e representa um documento por um *array* em que, para cada posição correspondente ao índice da palavra, é colocado um número, e as palavras mais frequentes ocupam as primeiras posições do *array*.

No projeto Cladop, foram experimentados 4 modos de conversão de texto em número (parâmetro *bag_ind_modo_numerico)*: *binary, count, tfidf* e *freq*. O modo definido foi usado tanto para substituir palavras do conteúdo quanto dos nomes dos documentos. As que levaram a uma maior acurácia foram *tfidf* e *binary*, nessa ordem, conforme pode ser observado no Gráfico 6.

Gráfico 6 – *Boxplot* do percentual de acurácia em base de teste por modo de conversão de palavra em número[50]

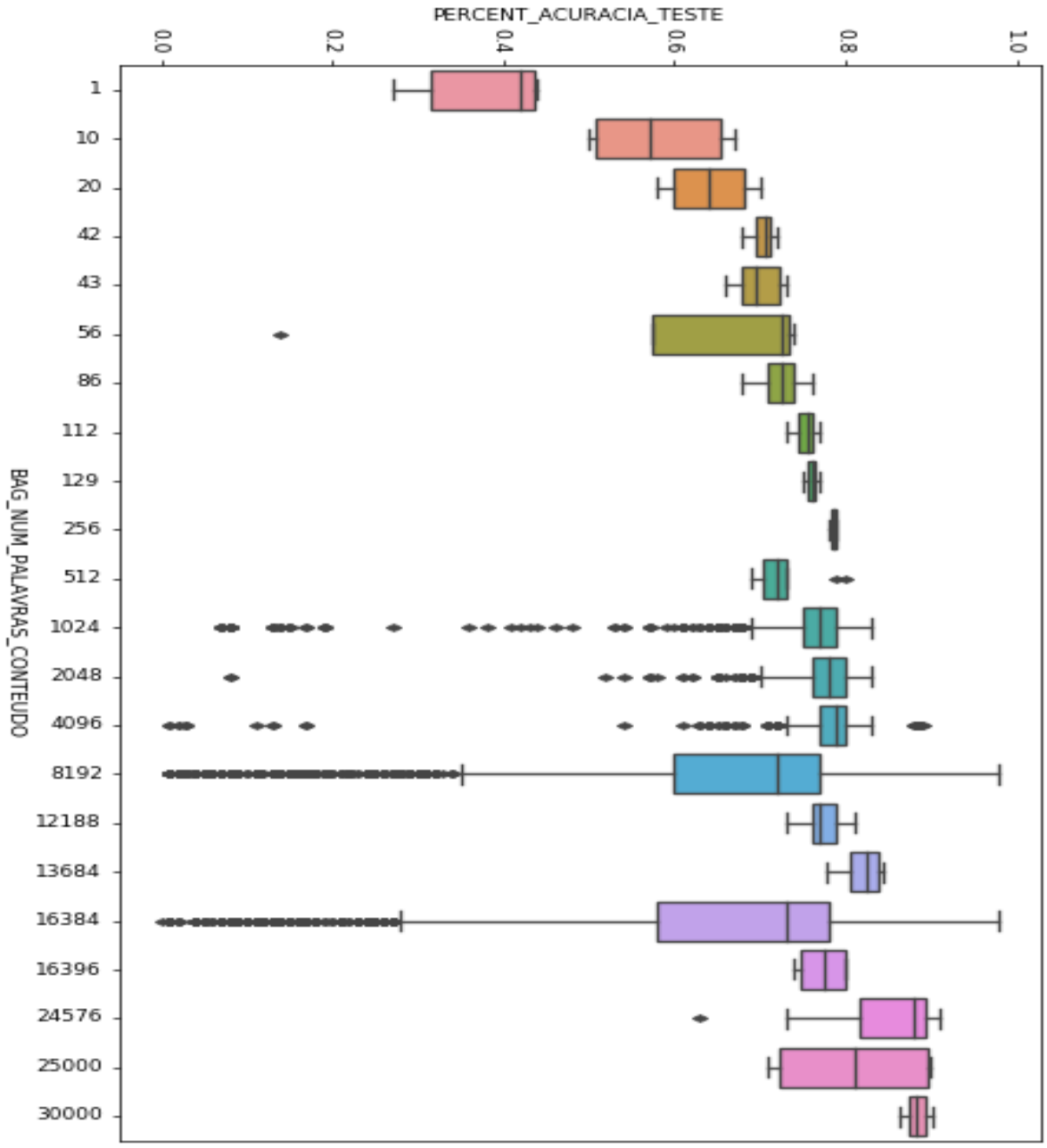


Fonte: Elaborada pelo autor (2019).

[50] Todos os *boxplots* do relatório representam os 12.866 treinamentos efetuados e dão uma noção da distribuição do percentual de acurácia em testes alcançados (eixo y) para cada valor de parâmetro (eixo x).

Os tamanhos dos *arrays* também são parâmetros dos treinamentos, denominados *bag_num_palavras_conteúdo* e *bag_num_palavras_nome_arquivo*, associados respectivamente ao conteúdo e ao nome arquivo. Conforme Gráfico 7, um tamanho maior de *bag* de conteúdo proporciona uma maior acurácia para o classificador.

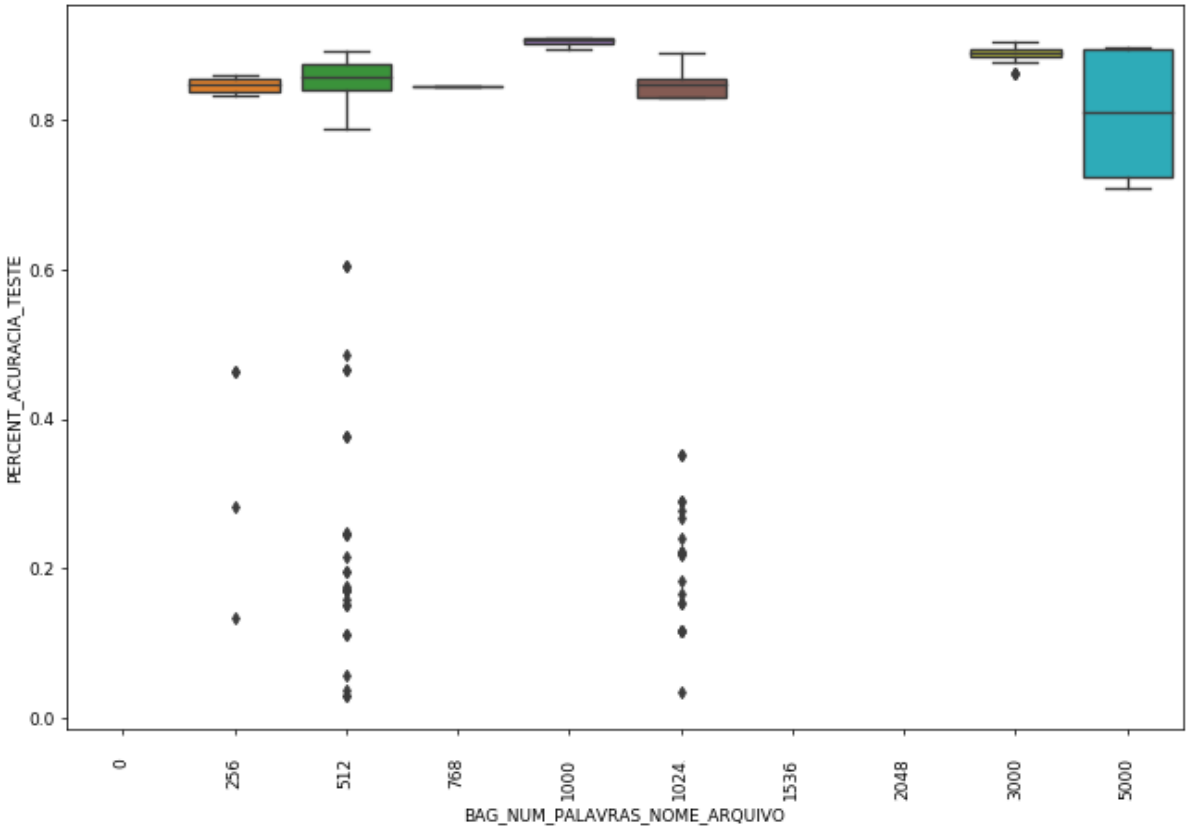
Gráfico 7 - *Boxplot* do percentual de acurácia em base de teste por tamanho do bag do conteúdo do documento



Fonte: Elaborada pelo autor (2019).

Por ser um contexto diferente, foi criado um *bag* à parte para nome do arquivo. Após a construção do *bag* para nomes de arquivos, algumas siglas e abreviações do negócio encontradas foram incorporadas ao dicionário de palavras válidas do negócio enriquecendo o *bag* do conteúdo. Foi observado que apenas as 1000 palavras mais comuns eram importantes para o classificador, conforme pode ser percebido no Gráfico 8.

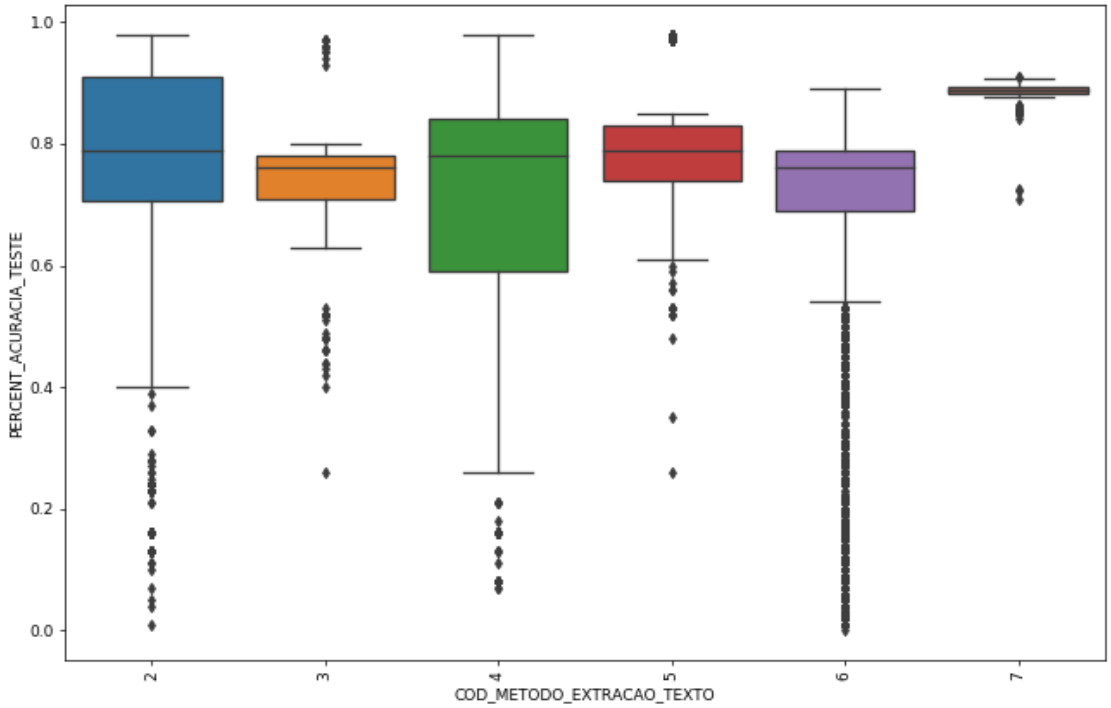
Gráfico 8 - *Boxplot* do percentual de acurácia em base de teste por tamanho do bag do nome do documento



Fonte: Elaborada pelo autor (2019).

O Gráfico 9 traz a acurácia alcançada pelos diferentes métodos aplicados de pré-processamento.

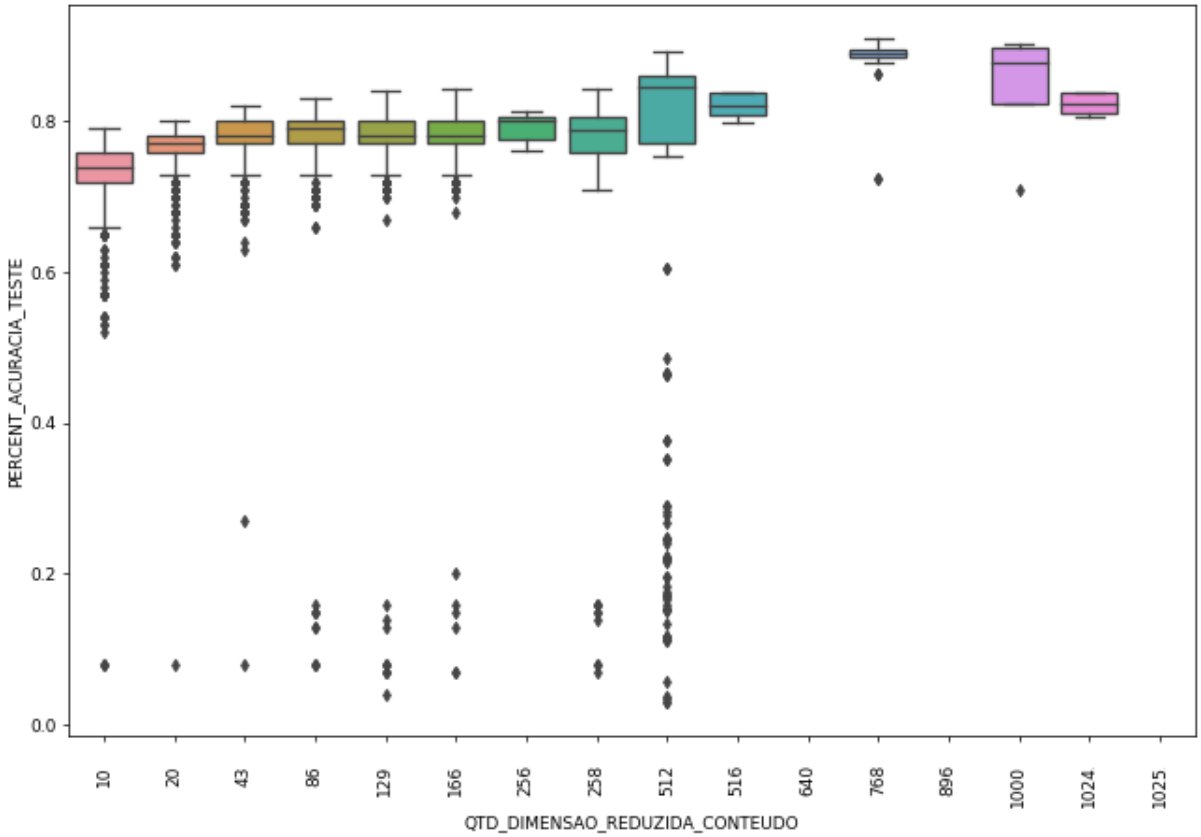
Gráfico 9 - *Boxplot* do percentual de acurácia em base de teste por método de pré-processamento



Fonte: Elaborada pelo autor (2019).

Para se evitar a *maldição da dimensionalidade*, que, resumidamente, alerta para o risco de um elevado número de varáveis impactar negativamente o modelo gerado, os treinamentos recebem como parâmetros, de forma opcional, o tamanho a ser derivado a partir do *bag* e a técnica a ser usada, denominados, respectivamente, *qtd_dimensao_reduzida_conteudo* e *redutor_dimensao_conteudo*. O Gráfico 10 ilustra o comportamento da acurácia do modelo em relação ao número de variáveis resultantes da redução de dimensões do *array,* independentemente da técnica usada.

Gráfico 10 - *Boxplot* do percentual de acurácia em base de teste por número de variáveis após redução dimensões do *array*



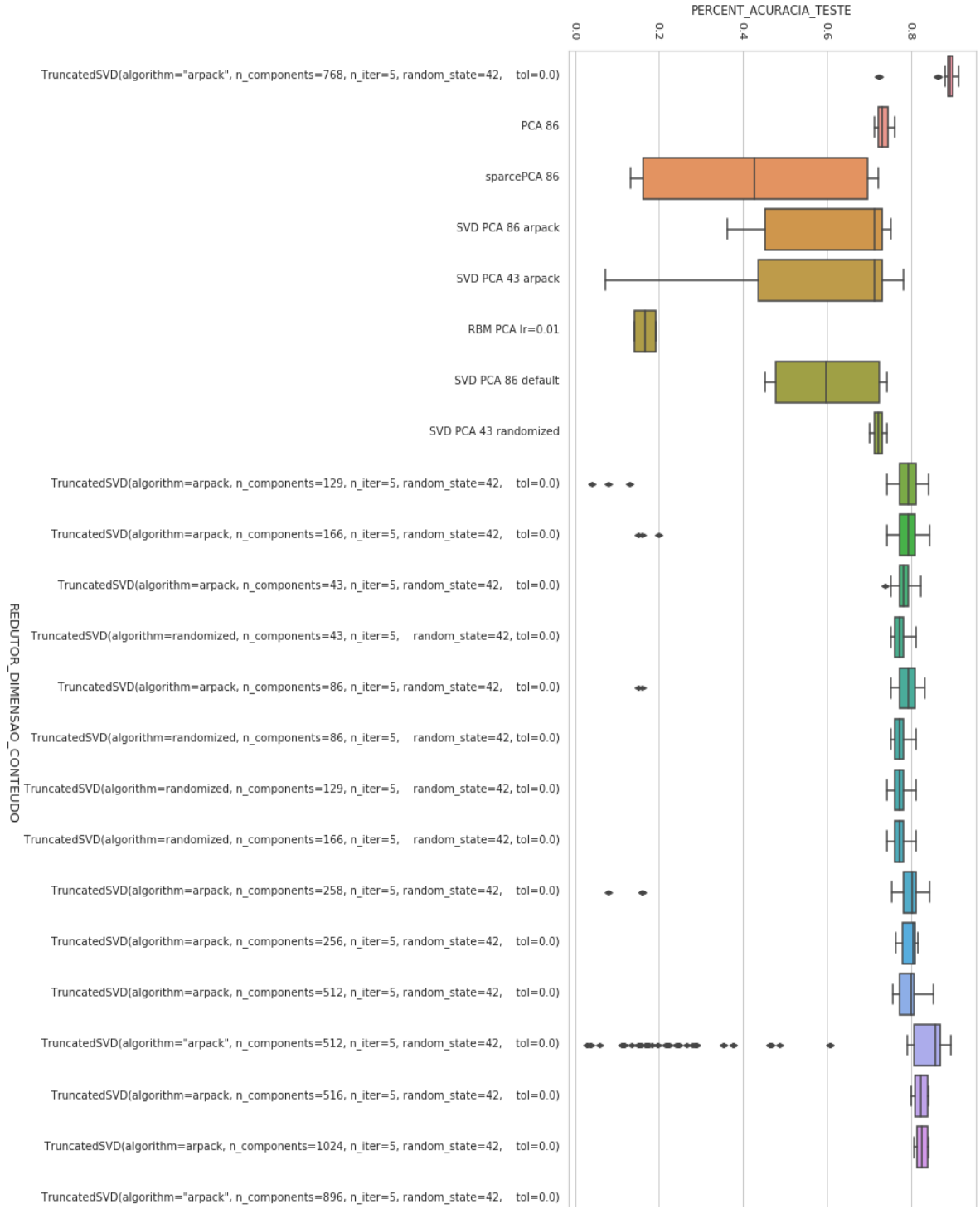
Fonte: Elaborada pelo autor (2019).

Algumas técnicas de redução foram experimentas. A que proporcionou a maior acurácia com o menor tempo de resposta foi *TruncatedSVD*. O parâmetro passado para o treinamento é um objeto *python* que implementa a transformação da biblioteca *sklearn.decomposition.* O Gráfico 11 demonstra a variação da acurácia para alguns redutores.

O Gráfico 12 traz uma matriz de dispersão[51] entre os parâmetros relacionados à troca de palavras por números.

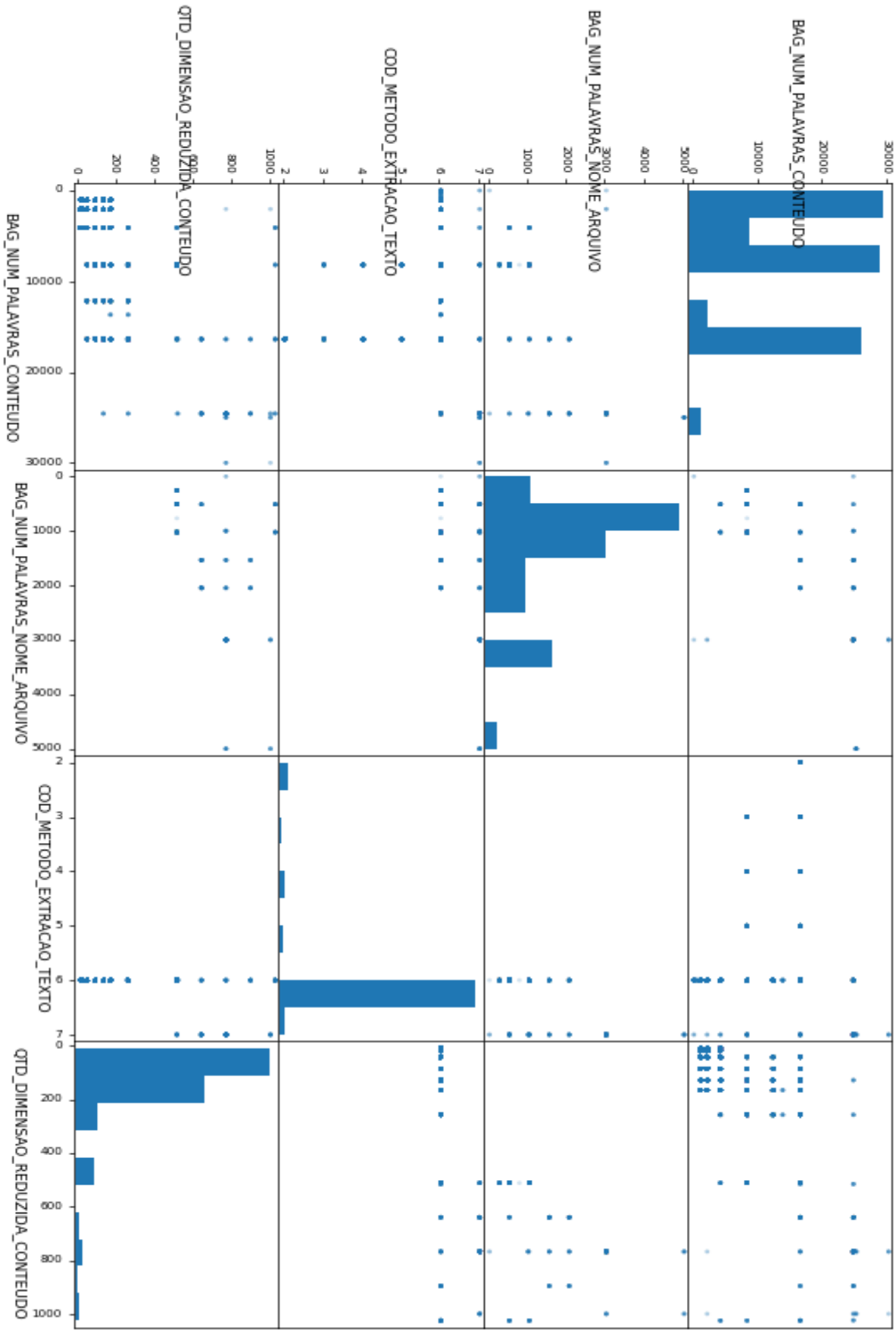
[51] Trata-se de uma matriz de gráficos de dispersão, que, segundo *wikipedia*, utilizam coordenadas cartesianas para exibir valores de duas variáveis. Os dados são exibidos como uma coleção de pontos, cada um com o valor de uma variável determinando a posição no eixo horizontal e o valor da outra variável determinando a posição no eixo vertical. No Gráfico 12, os pontos correspondem aos registros de treinamentos que usaram os valores indicados das duas variáveis e permitem visualizar os valores mais experimentados concomitantemente.

Gráfico 11 - *Boxplot* do percentual de acurácia em base de teste por redutor de dimensões aplicado - parciail



Fonte: Elaborada pelo autor (2019).

Gráfico 12 – Matriz de dispersão (*Scatter matrix)* entre os parâmetros relacionados à substituição de palavras por números



Fonte: Elaborada pelo autor (2019).

A Tabela 12 relaciona os registros de definições de ação e de aprendizados relacionados à atividade de substituição de palavras por números em ordem cronológica.

Tabela 12 - Registros de definições de ação e de aprendizados relacionados

Momento	Descrição
26/2/19 14:16	Aplicando ln nas variáveis de qualidade do documento e testando tamanho menor de bag
26/2/19 16:20	Tentar PCA para reduzir bag. Manter: ln nas variáveis de qualidade do documento
27/2/19 10:21	O uso das 129 primeiras palavras do bag trouxe um resultado (75%) pouco inferior ao uso de pca 129 (random) de 79%
27/2/19 10:29	Em MLP, o uso de tamanho bag pequeno (125) com as variveis extras trouxe um resultado (76%) pouco inferior ao uso somente de bag com muitas palavras (depende do otimizador). Logo, para um bag pequeno, as variveis extras fazem diferença.
27/2/19 10:34	Definindo o PCA a partir de limite de variância: Resultados com PCA (sklearn) kind bag tamanho treino limitevar n_components_, Acurácia(so palavras) full tfidf 4096 4808 0.8 522 0.6621 full tfidf 4096 4809 0.9 1018 0.67571 full tfidf 4096 4811 0.9 1999 0,67 full binary 4096 4812 0.9 1872 0,54 full tfidf 8192 4813? 0.9 3249 0.669
27/2/19 10:39	Ao usar pca (sklearn): melhor separar o comando fit do comando transform. O comando fit_transform estava travando!
27/2/19 10:53	Ao usar pca (sklearn): se o número de dimensões for bem pequeno e o array for grande (ver detalhes na documentação da função), melhor usar o método randomized do que o full. Pois é mais rápido, e os resultados (variância alcançada) são equivalentes.
1/3/19 18:09	Conclusões: (com variáveis extras; MLP com otimizador, adam, tfidf, etc): . usar pca com o mesmo número de classes se mostrou a opção mais benéfica (custo x benefício)! Talvez possamos pressupor a divisão dos textos em clusters de forma equivalente ao número de classes. . aplicar o pca no array com 1024 palavras mais comuns se mostrou mais benéfico do que aplicar o pcar no array com 8192 palavras. Não porque a variância explicado foi maior (59% contra 37%), isso era esperado de um array menor. Mas por causa da acurácia alcançada (78.4 x 78.06). . sem juntar as variáveis extras, o uso de PCA (n=86) não leva ao mesmo resultado do uso do bag completo (n=8192):77.06 x 80%. Mas, ao concatenar as variáveis, o resultado com PCA se aproximou mais (78.4% x 80%).
1/3/19 18:10	Testando só bag menor (sem pca) com variáveis extras bag mlp_adam randomforest 129 76 86 75,85 75.10 43 72.45 73.45 20 70 10 66.5 Testando só bag menor (sem pca) sem variáveis extras bag randomforest 86 71.54 43 68.33 20 62 10 50.80
1/3/19 18:14	Considerando randomforest (tipo sem outros): O bag menor (86) sem variáveis extras ja traz um resultado interessante: 71% Com variáveis extras, há um aumento para 75,10%. Mlp aumenta para 75,85%. As 10 palavras mais comuns sem variáveis extras levam a acurácia de 50.80%. Com variáveis extras: 66.5%
1/3/19 18:20	Outras formas de redução de dimensionalidade Bag 1024 reduzindo por variância (acurácia; randomforest) variáveis extras: sem com tamanho obs sparcePCA 72.19 86 muito demorado

	LDA 71,85 86 default RBM 18.6 86 lr=0.001 TruncatedSVD 74,14 75,41 86 (ardock) - super rápido TruncatedSVD 73,26 75,8 43 (ardock) - super rápido Melhor configuração: truncatedsvd: default, algorithm=ardock randomforest: default + {bootstrap: False, n_estimators: 20} acurácia alcançada em teste: 0.7326
1/3/19 18:22	Truncated SVD com tamanho 43 (metade do número de classes) se mostrou mais benéfico (execução rápida e melhor acurácia - com randomforest) do que SparcePCA (muito lento!) e RBM (tem baixa acurácia)
1/3/19 18:23	Experimentando TruncatedSVD (merece ser revista a execução) partindo de 8192 para 86 dimensões; sem variáveis extras .randomized, ficou: .. pior: 48.06% (n_iter=10 ou 100) .arpack, ficou: .. pior: 48.06% (n_iter=10) partindo de 8192 para 166 dimensões; sem variáveis extras .arpack (n_iter=10): 48.06 partindo de 1024 para 86 .arpack:74.07 (n_iter=100) ou 0.7 ???? partindo de 1024 para 43 dimensões; sem variáveis extras .arpack (n_iter=10): 74.34 .randomized (n_iter=10): 73.89 partindo de 1024 para 20 dimensões; sem variáveis extras .arpack (n_iter=10): 71.72 partindo de 1024 para 166 dimensões; sem variáveis extras .arpack (n_iter=10): 73.50 partindo de 512 para 43 dimensões; sem variáveis extras .arpack (n_iter=10): 72.93 ..randomized (n_iter=10): 72.68 partindo de 512 para 86 dimensões; sem variáveis extras .arpack (n_iter=10): 72.83 Melhor: partir de 1024; usar 43 dimensões e método arpack
15/3/19 17:38	Otimizador adadelta conseguiu melhor acurácia até o momento (84.4). Como sugerido pelo Eric, parece trabalhar bem em matriz esparsa de bag com 13684 e sem colunas extras. Tfidf: 84.4; binary: 84.3; colunas extras(258):84,2. Para bag menor, com 8192 palavras, e tSVD de 166, obteve também 84.2.
19/3/19 8:55	Avaliar possibilidades para tratamento de nomes de arquivos: retirar todos os números: retirar ""num ordem - "" Como em: 04 - Nota de Empenho.pdf Retirar números do texto 10.2 Formulários de prestação de contas1 - OCR.PDF 13.2.5 manifestação 3.5.PDF 65.4 Aviso de recebimento4.PDF 76.4. Parecer da Auditoria_2018_06_13_11_31_34_655.PDF 16.7 aviso de recebimento_2018_05_25_13_05_56_204.PDF 2. Estatuto da Beneficiaria (2).PDF sei_093890239

	Retirar acento e colocar lower portaria de aprovação Retirar separadores _ - ( ) e espaço (deixar apenas 1 espaço) relatorio_extrato_cc.pdf ata (publicação dou).pdf sei_0833.pdf 6o ta)nubyta,odf retirar preposições e artigos (stopwords?) termo de cv.pdf retirar letras sozinhas resultado c.pdf tratar apóstrofe e ponto ao final da palavra ar s fns parecer tec. e financ. avaliac. consulta cpf s ar para e.tce Tratar lista de sinônimos: tec por técnico; tecncica por técnico; (acrescentar nas revisoes) retirar ""tipo documento"" (as vezes tem .pdf.pdf) ".pdf"" Não retirar palavras desconhecidas (pelo menos as mais comuns... talvez pegar as 1as do bag) ar.pdf consulta sisgru cadin audit parcer etce consulta cpf s Estratégia: Montar um bag proprio Identificar sinônimos entre os mais frequentes (1000) retirar sinônimos tec por tecnico Avaliar possibilidade de acrescentar as palavras dos nomes na construção do bag maior de palavras
12/4/19 9:56	Criado modelo 7 de tratamento de texto com classes novas: nome, sobrenome, siglauf e nomeuf. Bem como integrado com as palavras mais comuns no modelo de nomes de arquivos. Também feita tradução de palavras comuns com erro: aenxo por anexo
18/4/19 11:28	Ensemble alcançou 90,2% para bags maiores e logisticregression como metaclassifier.
15/5/19 7:44	Para a rede 7, o binary demorou mais (36 contra 33 épocas) e teve acurácia um pouco pior. Tfidf: 89,9 +- 0,3 contra 89,5 +- 0.2.
22/5/19 16:47	Constatada (para LGBMClassifier) um aumento na acurácia em teste para uso de TFIDF ao invés de BINARY na casa de 0.4% em média para as últimas execuções, envolvendo mais documentos
5/6/19 18:20	Passamos a considerar para nome de arquivo um bag de 1000 palavras, pois constatamos pelo lgbmClassifier.booster_.feature_importance() que a partir de 1000 a importância era zero. E a *rede MLP* melhorou 0.15% sua acurácia.
10/6/19 18:57	Rodando novamente a execução de produção só para salvar o redutor (truncatedsvd) e apurar precisão por tipo em dados de validação

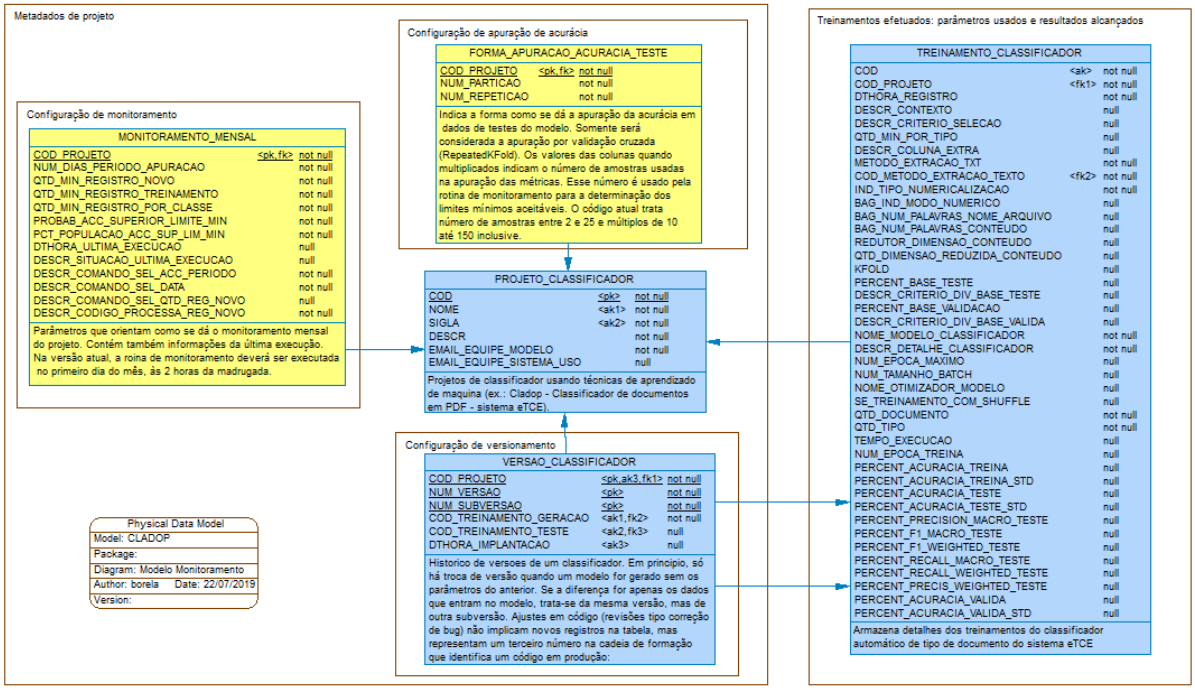
Fonte: Elaborada pelo autor (2019).

4.6 ACOPLAMENTO DO RASTRO COM ATIVIDADES DE MONITORAMENTO

A rotina de monitoramento da eficácia de um classificador em produção construída no projeto Cladop é um bom exemplo de benefício do rastro.

A Figura 10 apresenta um modelo Entidade-Relacionamento (MER) integrando novos metadados de monitoramento do projeto (tabelas em amarelo) aos conceitos já apresentados na Figura 8, em azul, referentes aos treinamentos do classificador.

Figura 10 - Modelo Entidade x Relacionamento - integrando monitoramento e rastro



Fonte: elaborada pelo autor (2019)

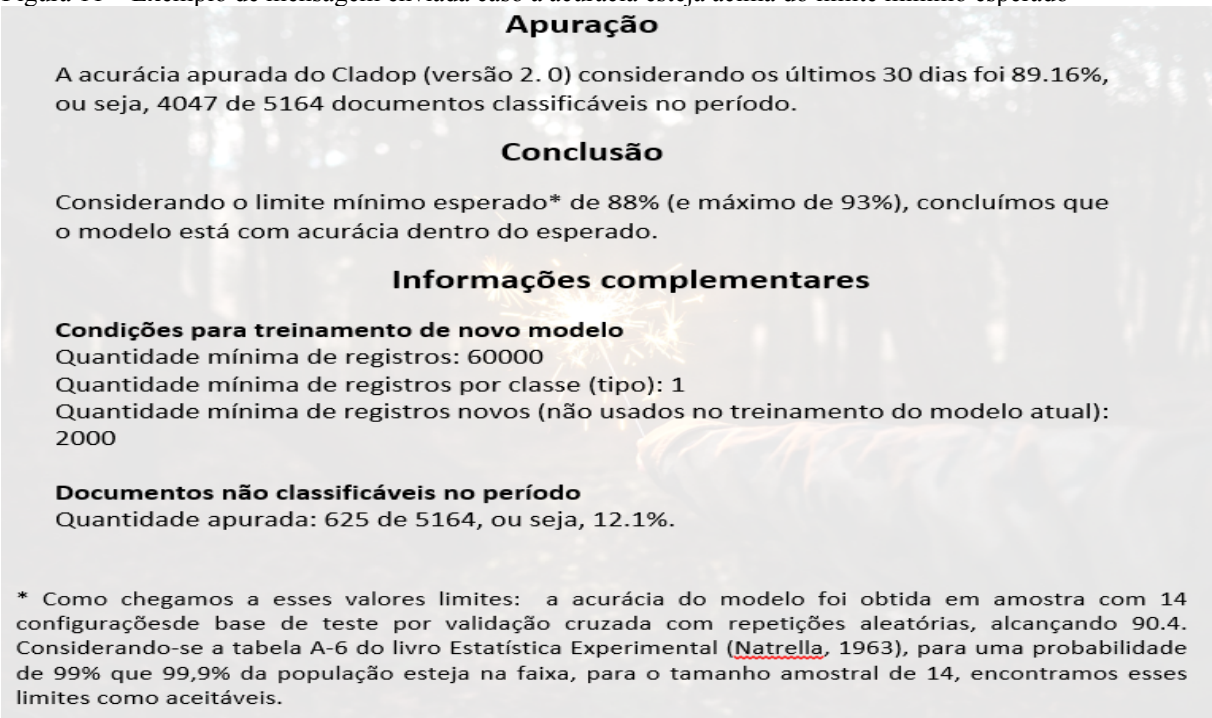
Como visto, a tabela *versao_classificador* traz informações do histórico de versões e subversões implantadas de um modelo. Para cada modelo, registram-se dois códigos de treinos do modelo: o treino em que se apurou a acurácia do modelo (*cod_treinamento_teste*) e o treino em que se gerou a versão final do modelo (*cod_treinamento_geracao*). O que diferencia uma versão de uma subversão são os critérios usados na sua construção. Se a mudança entre os modelos for apenas o conjunto de dados considerado no treinamento, tem-se uma nova subversão. Se a mudança for além dos dados envolvidos, por exemplo, implicando novos valores de hiperparâmetros ou o uso de outros tipos de algoritmos, tem-se uma nova versão.

A partir das informações do treino de teste, pode-se obter a acurácia alcançada do modelo e a forma de sua obtenção, indicada pelo número de partições e de repetições usadas.

Essas informações servem de base para a derivação da acurácia mínima aceitável para o modelo[52].

Contrasta-se a acurácia mínima aceitável com a acurácia efetiva apurada em dados reais de uso do classificador em um período, delimitado pela coluna *num_dias_periodo_apuracao* da tabela *monitoramento_mensal*. Essa acurácia deve ser obtida a partir do comando *sql* registrado na coluna *descr_comando_sel_acc_periodo* da mesma tabela. Se a acurácia apurada estiver acima do limite mínimo esperado, um e-mail é enviado à equipe do modelo avisando o resultado. A Figura 11 ilustra mensagem real enviada pela rotina proposta em uso para o Cladop.

Figura 11 – Exemplo de mensagem enviada caso a acurácia esteja acima do limite mínimo esperado



Apuração

A acurácia apurada do Cladop (versão 2. 0) considerando os últimos 30 dias foi 89.16%, ou seja, 4047 de 5164 documentos classificáveis no período.

Conclusão

Considerando o limite mínimo esperado* de 88% (e máximo de 93%), concluímos que o modelo está com acurácia dentro do esperado.

Informações complementares

Condições para treinamento de novo modelo
Quantidade mínima de registros: 60000
Quantidade mínima de registros por classe (tipo): 1
Quantidade mínima de registros novos (não usados no treinamento do modelo atual): 2000

Documentos não classificáveis no período
Quantidade apurada: 625 de 5164, ou seja, 12.1%.

* Como chegamos a esses valores limites: a acurácia do modelo foi obtida em amostra com 14 configuraçõesde base de teste por validação cruzada com repetições aleatórias, alcançando 90.4. Considerando-se a tabela A-6 do livro Estatística Experimental (Natrella, 1963), para uma probabilidade de 99% que 99,9% da população esteja na faixa, para o tamanho amostral de 14, encontramos esses limites como aceitáveis.

Fonte: Elaborada pelo autor (2019).

Contudo, se a acurácia for inferior ao limite mínimo, deve-se iniciar automaticamente um treinamento de um novo modelo com dados não usados no treinamento da versão atual em produção. Mas, se não houver um mínimo de registros novos, o treinamento deve ser suspenso e a equipe deve ser notificada dessa situação por e-mail. O limite mínimo de registros novos consta na coluna *qtd_min_registro_novo* e o comando para a identificação da quantidade de registros está em *descr_comando_sel_qtd_reg_novo*.

[52] O tamanho da amostra, a métrica apurada e o desvio padrão são pesquisados na Tabela A-6 do livro Estatística Experimental (Natrella, 1963) para obtenção dos limites aceitáveis. Os parâmetros de probabilidade e o percentual da população a serem considerados são obtidos, respectivamente, das colunas *probab_acc_superior_limite_min* e *pct_populacao_acc_sup_lim_min* da tabela *monitoramento_mensal*.

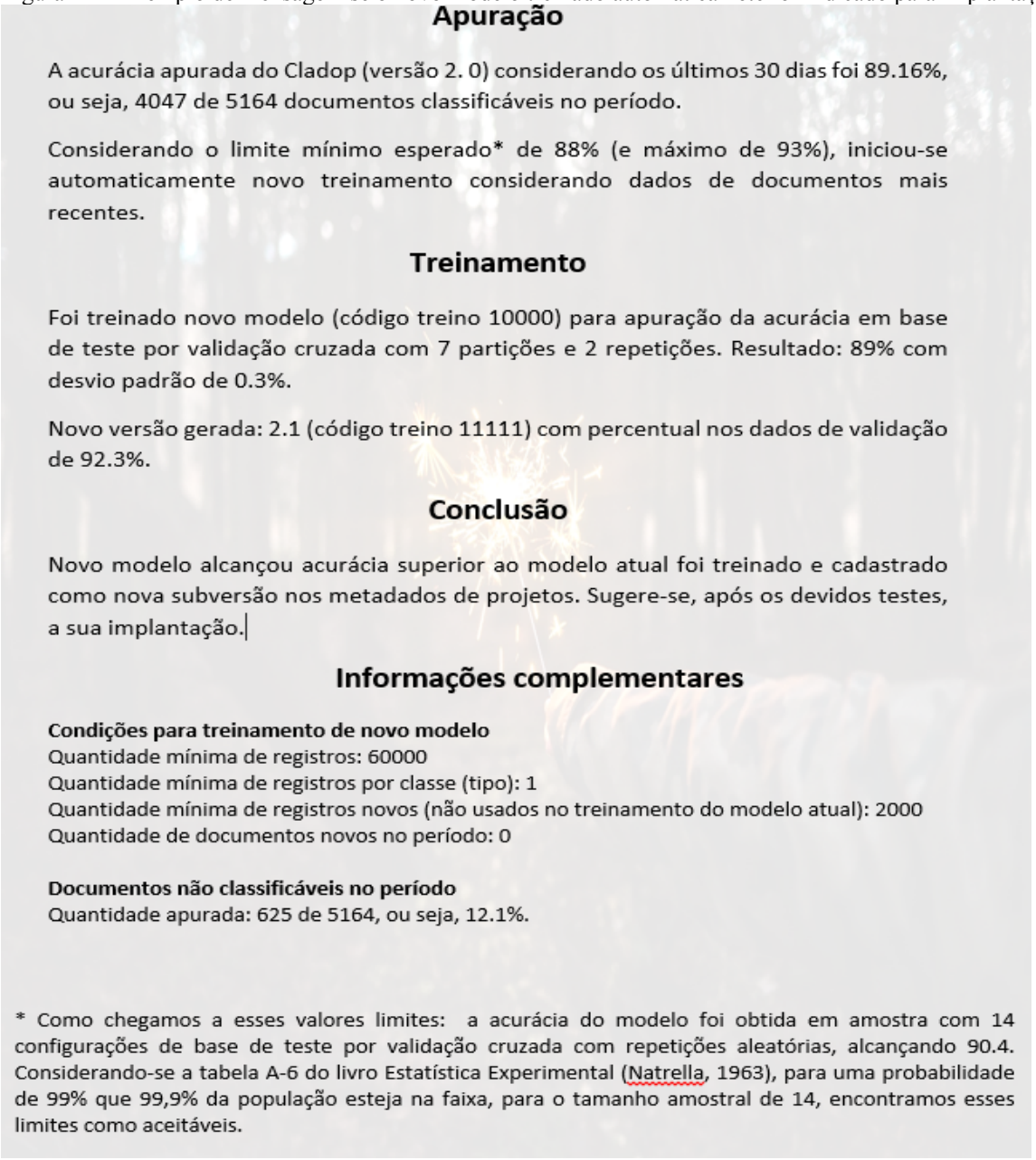
No caso de haver registros novos em número acima do parâmetro indicado, apura-se a acurácia do modelo diante dos novos dados, atendidos os demais requisitos mínimos indicados (*qtd_min_registro_treinamento* e *qtd_min_registro_por_classe*). A forma de apuração da acurácia se dá com validação cruzada seguindo o número de partições e de repetições indicadas na tabela *forma_apuracao_acuracia_teste*.

Antes do treinamento, pode-se, se necessário, efetuar o processamento de registros novos, através do código *python* indicado na coluna *descr_codigo_processa_reg_novo*. No caso do Cladop, a base espelho com os dados dos documentos disponíveis para uso no treinamento é atualizada e os textos pré-processados.

Se a acurácia obtida não superar a do modelo em produção, um e-mail é enviado à equipe do projeto para, se for o caso, rever os parâmetros em uso. Mas, se a acurácia for maior, parte-se para o treinamento automático do modelo final, usando-se os mesmos parâmetros da versão em produção, obtidos da tabela de treinamentos do rastro usando como chave a informação *cod_treinamento_geracao* da tabela *versao_classificador*. O novo treinamento é registrado no rastro e novo registro de subversão[53] é inserido na tabela de versões do modelo com os respectivos códigos de treinamento, e um e-mail deve ser enviado para a equipe ultimar os testes e avaliar a implantação do modelo treinado, como ilustrado na Figura 12.

[53] Trata-se de uma nova subversão do modelo, haja vista que a diferença foi apenas os dados considerados no treinamento do modelo, tendo sido mantidos os parâmetros indicados no treinamento da versão atual.

Figura 12 – Exemplo de mensagem se o novo modelo treinado automaticamete for indicado para implantação.



Apuração

A acurácia apurada do Cladop (versão 2. 0) considerando os últimos 30 dias foi 89.16%, ou seja, 4047 de 5164 documentos classificáveis no período.

Considerando o limite mínimo esperado* de 88% (e máximo de 93%), iniciou-se automaticamente novo treinamento considerando dados de documentos mais recentes.

Treinamento

Foi treinado novo modelo (código treino 10000) para apuração da acurácia em base de teste por validação cruzada com 7 partições e 2 repetições. Resultado: 89% com desvio padrão de 0.3%.

Novo versão gerada: 2.1 (código treino 11111) com percentual nos dados de validação de 92.3%.

Conclusão

Novo modelo alcançou acurácia superior ao modelo atual foi treinado e cadastrado como nova subversão nos metadados de projetos. Sugere-se, após os devidos testes, a sua implantação.

Informações complementares

Condições para treinamento de novo modelo
Quantidade mínima de registros: 60000
Quantidade mínima de registros por classe (tipo): 1
Quantidade mínima de registros novos (não usados no treinamento do modelo atual): 2000
Quantidade de documentos novos no período: 0

Documentos não classificáveis no período
Quantidade apurada: 625 de 5164, ou seja, 12.1%.

* Como chegamos a esses valores limites: a acurácia do modelo foi obtida em amostra com 14 configurações de base de teste por validação cruzada com repetições aleatórias, alcançando 90.4. Considerando-se a tabela A-6 do livro Estatística Experimental (Natrella, 1963), para uma probabilidade de 99% que 99,9% da população esteja na faixa, para o tamanho amostral de 14, encontramos esses limites como aceitáveis.

Fonte: Elaborada pelo autor (2019)

.

Ao final do processamento, atualiza-se nos metadados de monitoramento do projeto com o resultado alcançado e a data e a hora de término do treinamento (colunas *dthora_ultima_execucao* e *descr_situacao_ultima_execucao*).

Embora construída para atender o Cladop, qualquer outro classificador que use a mesma estrutura de rastro de treinamentos também pode ser monitorado[54]. Mais detalhes sobre o fluxo de processamento da rotina de monitoramento encontra-se no Apêndice C.

[54] A manutenção exigirá uma pequena alteração na rotina que hoje só espera um projeto na tabela *projeto_classificador*. Nada que um comando de *loop* não resolva facilmente.

Durante os testes da rotina de monitoramento, implementada pelo caderno *Cladop_monitoramento.ipynb*, foi treinado novo modelo com dados mais recentes, e foi detectado um aumento na acurácia, micro, de 90.4% para 91,1%, o que ensejou o registro da versão 2.1 do Cladop na tabela *versao_classificador* ainda a ser implantada[55].

5 CONCLUSÃO

Os objetivos propostos para o trabalho foram alcançados.

Uma breve contextualização com referencial teórico foi apresentada sobre metodologias e documentação em projetos de mineração de dados, bem como do potencial impacto das experiências adquiridas nos projetos para uma organização.

Foi proposto o Rastro-DM, que, em última instância, objetiva a retenção do conhecimento, fruto de experiências, de um projeto de DM. É um conjunto de boas práticas que, entre outras características apresentadas, é flexível e pode ser mesclado à metodologia em uso por uma organização.

Rastro-DM mostrou-se viável com a ilustração de sua aplicação no projeto Cladop, em que o rastro construído alcançou uma capacidade funcional suficiente para as necessidades do projeto. Foram mostrados como benefícios, entre outros, um relatório do projeto enriquecido com informações geradas automaticamente do rastro e uma rotina automática de monitoramento do classificador integrada ao rastro. O *kit Rastro-Dm*[56] do projeto Cladop encontra-se disponível para trabalhos futuros que se interessem em dar prosseguimento à reflexão aqui iniciada.

Há muito ainda por fazer. Trabalhos futuros podem experimentar o uso Rastro-DM em outros contextos: outras tarefas de mineração que não classificação, outras plataformas de desenvolvimento ou mesmo em outras culturas organizacionais. Rastro-DM pode ser revisado com o acréscimo de novas atividades ou mesmo enriquecido com técnicas e procedimentos que complementem as atividades sugeridas. Também é um desafio para trabalhos futuros a mineração de aprendizados a partir de rastros de treinamentos. E, talvez o mais importante

[55] Na Tabela 7 da seção 4.4.1 que apresenta os detalhes dos treinamentos de teste e de geração da versão 2.1 do Cladop, percebe-se o contexto de criação dos modelos: *Cladop_monitoramento.ipynb,* prova da integração entre o monitoramento e o rastro.

[56] *Kit Rastro-DM* é o nome dado pelo autor ao conjunto de arquivos (planilhas, por exemplo) com os dados das entidades rastreadas: treinamentos, definições de ação e aprendizados. Como dito, os dados do rastro no Cladop e o código usado para sua construção encontram-se publicados em https://gitlab.com/MarcusBorela/rastro-dm.git, na pasta Rastro_Projeto_Cladop.

ponto a evoluir, que não foi escopo deste trabalho, é a partilha das experiências adquiridas e o uso de táticas institucionais para o crescimento do conhecimento organizacional, pois preocupa o fato de a integração entre DM e KMP possuir um grande vazio de pesquisas (NGUYEN, 2018). Afinal, o valor dos dados está em como eles são interpretados e usados (BERMAN *et al*, 2018).

Por fim, ficou claro que, além de contribuir para a equipe do projeto, o rastro dos projetos tem potencial para promover um salto no conhecimento organizacional, se forem adotadas medidas institucionais para incentivar sua construção, sua partilha e a busca automática de aprendizados neles escondidos. Mas há que se ter em mente que nenhum direcionamento corporativo deve inibir a liberdade dos pesquisadores e dos analistas de dados, pois diminuiria sua criatividade. E a liberdade criativa, de construir o próprio rastro, é algo de que os seres humanos não podem prescindir, pois é um dos grandes diferenciais que os impedem de serem classificados como máquinas.

REFERÊNCIAS

APPEL-NASA-GOV. Knowledge Management Process. Appel-Nasa-Gov. Disponível em: https://appel.nasa.gov/wp-content/uploads/2015/11/Knowledge-Management-Process.pdf Acesso em 27 jul. 2019**.** 2015.

BECKER, Karin; GHEDINI, Cinara. A documentation infrastructure for the management of data mining projects. **Information and Software Technology**, v. 47, n. 2, p. 95-111, 2005.

BERMAN, Francine et al. Realizing the potential of data science. *Communications of the ACM*, v. 61, n. 4, p. 67-72, 2018.

BHATT, Ganesh D. Knowledge management in organizations: examining the interaction between technologies, techniques, and people. *Journal of knowledge management*, v. 5, n. 1, p. 68-75, 2001.

BRASIL. Constituição da República Federativa do Brasil. 1988.

BRASIL. Tribunal de Contas da União. Instrução Normativa nº 71, de 28 de novembro de 2012. Dispõe sobre a instauração, a organização e o encaminhamento ao Tribunal de Contas da União dos processos de tomada de contas especial. 2012.

_____. Decisão Normativa nº155, de 23 de novembro de 2016. Regulamenta os incisos I, III, IV, V e VI do art. 17 da Instrução Normativa - TCU nº 71, de 28 de novembro de 2012. 2016.

_____. Portaria nº122, de 20 de abril de 2018. Dispõe sobre a implantação e a operacionalização do sistema informatizado de tomada de contas especial (Sistema e-TCE), com amparo no § 5º do art. 11 da Decisão Normativa TCU nº 155, de 23 de novembro de 2016. 2018.

CASTANEDA, Delio Ignacio; MANRIQUE, Luisa Fernanda; CUELLAR, Sergio. Is organizational learning being absorbed by knowledge management? A systematic review. *Journal of Knowledge Management*, v. 22, n. 2, p. 299-325. 2018.

CHAPMAN, Pete et al. CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc*, v. 16, 2000.

CHOLLET, Francois. **Deep Learning with Python**. Manning Publications Co., Greenwich, CT, USA. 2017.

CONKLIN, Jeffret. Capturing Organisational Memory. In: Groupware and Computer-Supported Cooperative Work, R.M. Barcker (Ed.), Morgan Kaufman, pp. 561-565. 1996.

DINGSØYR, Torgeir; Moe, Nils Brede; Øystein. Nytrø. Augmenting experience reports with lightweight postmortem reviews. *Lecture Notes in Computer Science*, 2188:167–181, 2001.

GHEDINI, Cinara; BECKER, Karin. KDD application management through documentation. Disponível em: https://www.researchgate.net/profile/Karin_Becker2/publication/268253354_KDD_application_management_through_documentation/links/5657a5ec08ae1ef9297bf1d1/KDD-application-management-through-documentation.pdf. Acesso em 27 jul. 2019**.** 2000.

_____. A documentation model for KDD application management support. In: SCCC 2001. 21st International Conference of the Chilean Computer Science Society. IEEE. p. 105-114. 2001.

GREFF, Klaus et al. The sacred infrastructure for computational research. In: *Proceedings of the Python in Science Conferences-SciPy Conferences*. 2017.

HUBER, Steffen et al. DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. *Procedia CIRP*, v. 79, p. 403-408, 2019.

KURGAN, Lukasz A.; MUSILEK, Petr. A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, v. 21, n. 1, p. 1-24, 2006.

MARBÁN, Óscar et al. An engineering approach to data mining projects. In: *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, Berlin, Heidelberg, p. 578-588. 2007.

MARISCAL, Gonzalo; MARBAN, Oscar; FERNANDEZ, Covadonga. A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, v. 25, n. 2, p. 137-166, 2010.

MINGERS, John; BROCKLESBY, John. Multimethodology: Towards a framework for mixing methodologies. *Omega*, v. 25, n. 5, p. 489-509, 1997.

NATRELLA, M. G. **Estatística Experimental**., NBS Handbook 91. 1963.

NGUYEN, Ngoc Buu Cat. Data Mining in Knowledge Management Processes: Developing an Implementing Framework. 2018.

PRAKASH, BV Ajay; ASHOKA, D. V.; ARADHYA, VN Manjunath. Application of data mining techniques for software reuse process. *Procedia Technology*, v. 4, p. 384-389, 2012.

PUBLIO, Gustavo Correa et al. ML-Schema: Exposing the Semantics of Machine Learning with Schemas and Ontologies. *arXiv preprint arXiv:1807.05351*, 2018.

STATA, Ray. Organizational learning: The key to management innovation. Massachusetts Institute of Technology, 1980.

W3C (World Wide Web Consortium) Machine Learning Schema Community Group. W3c machine learning schema. Disponível em: https://www.w3.org/community/ml-schema. Acesso em 30 jul. 2019. 2017.

WIRTH, Rüdiger; HIPP, Jochen. CRISP-DM: Towards a standard process model for data mining. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Citeseer, p. 29-39. 2000.

APÊNDICE A – Exemplo de rastro persistido em arquivos locais

O código que se seque objetiva ilustrar uma implementação simples de uma infraestrutura de construção de um rastro em *python*. Ele fez parte da construção de um classificador[57]. Os dados de treinamentos (que engloba as definições ações que contextualizam o treinamento) e dos aprendizados são mantidos em *df_avaliacao* (*dataframe pandas*) e em *lista_obs* (*python list*) respectivamente. Eles são persistidos em pasta local em arquivos *.csv* e *.pickle* respectivamente.

O código não objetiva estar completo (inclusive faltam os comandos de importação das bibliotecas usadas) ou mesmo sem erros, mas ser apenas uma demonstração de que um rastro simples pode ser criado a um baixo custo com arquivos locais.

Versão mais completa deste código com dados do rastro local gerado encontra-se publicada em https://gitlab.com/MarcusBorela/rastro-dm.git, na pasta Rastro_Local.

Código de funções de apoio

```
def classifica(df_avaliacao, classificador, obs, X, y, kf,  verbose = True):
   t_start = time.time()
   cv_results = cross_validate(classificador, X, y, cv=kf,
         scoring=('accuracy', 'precision_weighted', 'f1_weighted', 'recall_weighted' ),
         return_train_score=True,
              n_jobs=-1)
   t_end = time.time()
   t_diff = t_end - t_start
   df_avaliacao = add_df_avaliacao(df_avaliacao, obs, cv_results, classificador, kf, len(y))
   if verbose:
      print("tempo: ", round(t_diff,3), " resultado:" ,list(df_avaliacao.iloc[-1][["test_accuracy_mean","train_accuracy_mean","classificador"]]))
```

[57] Trata-se de um código que é parte de um caderno *ipython* referente a um trabalho de uma disciplina da pós-graduação que deu origem a este Trabalho. Em algumas atividades de disciplinas da pós-graduação, o autor optou por codificar um rastro local de forma a viabilizar, ao final do processo, a geração automática dos aprendizados e das execuções implementadas, enriquecendo assim o resultado da missão.

```
    return df_avaliacao

def add_df_avaliacao(df, obs, cv_results, classifier, kf, tamanho):
    list_value=[obs, tamanho, str(type(classifier))[8:-2], str(classifier.get_params()), str(kf)]
    for key in cv_results:
        list_value.append(cv_results[key].mean())
        list_value.append(cv_results[key].std())
    return df.append(pd.DataFrame(np.array(list_value).reshape((1,25)), columns=list(df.columns)), ignore_index=True)

def apagar_ultimas_n_execucoes(qtd):
    global df_avaliacao
    df_avaliacao = df_avaliacao.iloc[:-qtd]
    return

def classifica_lista(df_avaliacao, obs, X, y, kf, lista_classificadores, verbose = True):
    for classifier in lista_classificadores:
        t_start = time.time()
        df_avaliacao = classifica(df_avaliacao, classifier, obs, X, y, kf,  verbose = False)
        t_end = time.time()
        t_diff = t_end - t_start
        if verbose:
            print("tempo: ",round(t_diff,3)," resultado:" , list(df_avaliacao.iloc[-1][["test_accuracy_mean",
                  "train_accuracy_mean","classificador"]]))
    return df_avaliacao

def cria_df_avaliacao():
    list_label=["obs", "tamanho", "classificador",   "parametros_clf" , "kfold"]
    list_value=[" ","", "" , "", ""]
    cv_results = cross_validate(RandomForestClassifier(n_estimators=80,random_state=vrandom_state), X, y, cv=kf,
              scoring=('accuracy', 'precision_weighted', 'f1_weighted', 'recall_weighted' ), return_train_score=True, n_jobs=-1)
    for key in cv_results:
        list_label.append(key+"_mean")
```

```
        list_label.append(key+"_std")
        list_value.append(cv_results[key].mean())
        list_value.append(cv_results[key].std())
    avaliacao = pd.DataFrame(np.array(list_value).reshape((1,25)), columns=list_label)
    avaliacao.drop([0], inplace=True)
    return avaliacao


def salva_ambiente():
    with open('observacoes.pickle', 'wb') as f:
        pickle.dump(lista_obs, f)
    df_avaliacao.to_csv("execucoes.csv")
    return

def recupera_ambiente():
    with open('observacoes.pickle', 'rb') as f:
        lista_obs=pickle.load(f)
    df_avaliacao = pd.read_csv("execucoes.csv", index_col=0)
    return lista_obs, df_avaliacao

def imprime_lista_obs():
    for item in lista_obs:
        print(" "); print(item)
    return

def apagar_ultimas_n_obs(qtd):
    global lista_obs
    lista_obs = lista_obs[:-qtd]
    return
```

Código ilustrando o uso das funções

```
# sintetizando aprendizados
        lista_obs.append("Experimentamos retirar as colunas de CEP (kfold 10), mas o resultado piorou: LRegression de 67.88 para 63.03; RForest
de 67.36 para 63.32 e MLP de 67.06 para 63.41. ")
        lista_obs.append("Com o stacking, a acurácia alcançada foi de surpreendentes 81.6% (ainda não usado kfold; base de teste com 12,3%)")

# definindo ação
# pode ser registrada de forma desvinculada de treinamentos, como observações
        lista_obs.append("Próximos passos: gridsearch no RandomForest e montar um stacking")

# ou vinculada a treinamentos, como parâmetro do treinamento
        descr_ação="Base completa; Dimensões reduzidas de 221 para 100"
        df_avaliacao = classifica_lista(df_avaliacao, descr_ação, X_dim_red, y, kf, lista_classificadores, True)

# executando experimentações
        kf = Kfold(n_splits=10, random_state=vrandom_state, shuffle=True)
        X, y = # foge ao escopo deste trabalho o detalhe dos dados

# um único treinamento de um classificador
        clf = DecisionTreeClassifier()
        df_avaliacao = classifica(df_avaliacao, clf, descr_ação, X, y, kf,  True)

# experimentando vários classificadores
        lista_classificadores = [ LogisticRegression(random_state=vrandom_state),
            RandomForestClassifier(n_estimators=80,random_state=vrandom_state),
            MLPClassifier(alpha = 1,random_state=vrandom_state)]
        obs="Base sem CEP, kfold 10, dimensoes reduzidas de 122 para "+ str(redutor.n_components)+" por " + str(redutor)
        df_avaliacao = classifica_lista(df_avaliacao, obs, X_dim_red, y, kf, lista_classificadores, True)

# imprimindo lista de aprendizados
        imprime_lista_obs()
```

```
# imprimento lista de treinamentos e ações motivadoras
        print(df_avaliacao)

# persistindo o rastro em arquivos locais
        salva_ambiente():

# recuperando rastro dos arquivos locais
        lista_obs, df_avaliacao = recupera_ambiente():
```

Alguns registros de aprendizados

- No contexto experimentado, a retirada de 150 outliers piorou a acurácia, passando de 70.74 para 70.68%. Outliers identificados por LocalOutlierFactor(n_neighbors=20, contamination=0.015).

- Ao reduzir as dimensões da base completa para 100 dimensões, a acurácia aumentou de 65.17% para 67.26.

- Com *kfold*:10 (base completa; tSVD 221:100), os algoritmos melhoraram um pouco na acurácia em relação a *kfold*:5, pois havia mais dados para treinamento (90%).

- Voltando todas as dimensões e usando *kfold*:10 (base completa): o LRegression aumentou de 66.91 para 67.88 e o MLP subiu muito pouco: de 66.80 para 67.06. O RForest (usado como contexto para validar o uso de redutor) reduziu de 67.36 para 65.19.

- Experimentamos retirar as colunas de CEP (*kfold* 10), mas o resultado piorou: LRegression de 67.88 para 63.03; RForest de 67.36 para 63.32 e MLP de 67.06 para 63.41.

- A redução de dimensões (tSVD-42 ou PCA-0.9) na base sem colunas de cep diminuiram um pouco (0.5 a 1%) a acurácia dos classificadores.

- Com o grid search com outros parâmetros, a acurácia reduziu para 65.93%. Faltou fazer o grig search com dimensões reduzidas.

- A melhor acurácia alcançada foi com o Stacking (81.6%)

Alguns registros de treinamentos com definições de ação de contexto (campo obs)

Tabela 13 - Alguns registros de treinamentos (visão parcial das colunas)

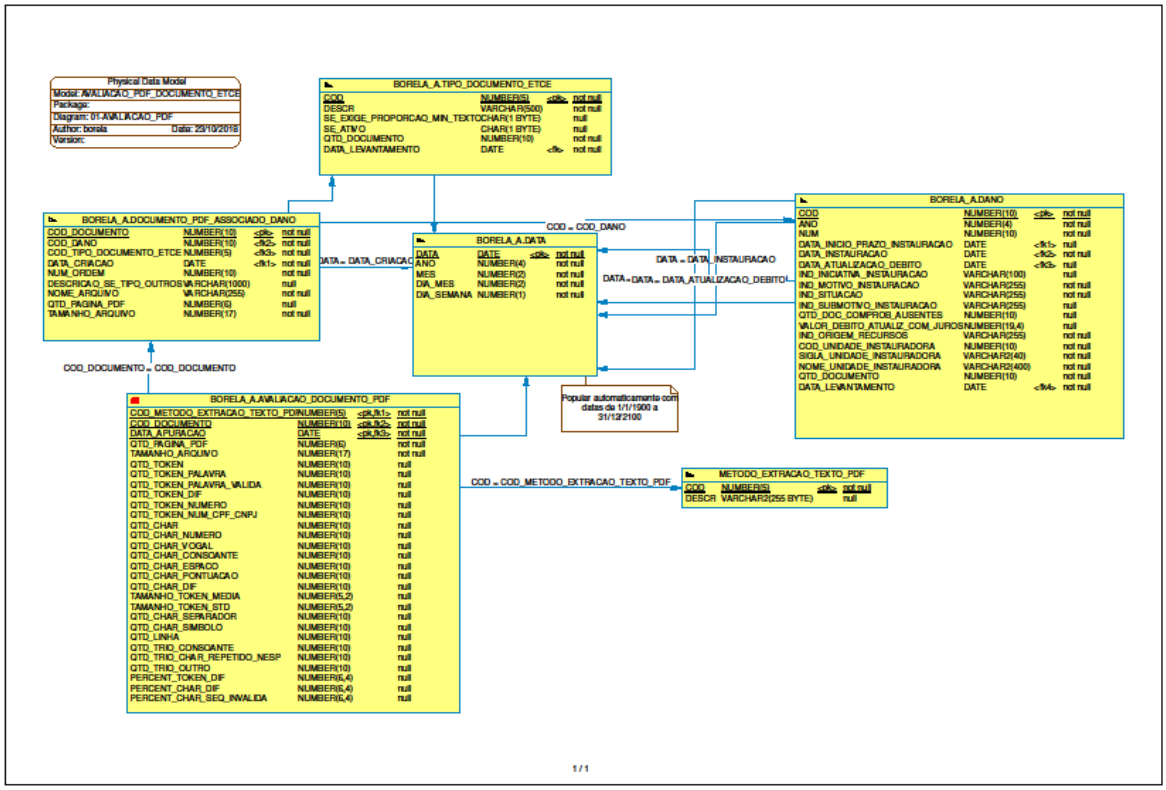
num	Obs (Definição de ação)	tamanho	modelo	Parâmetros classificador	Kfold	Fit time mean	Test neg mean absolute error mean	Test r2 mean
46	Base completa; X com 50 dimensões reduzidas por TruncatedSVD(algorithm='randomized', n_components=50, n_iter=5, random_state=None, tol=0.0)	38767	sklearn.linear_model.bayes.BayesianRidge	{'alpha_1': 1e-06, 'alpha_2': 1e-06, 'compute_score': False, 'copy_X': True, 'fit_intercept': True, 'lambda_1': 1e-06, 'lambda_2': 1e-06, 'n_iter': 300, 'normalize': False, 'tol': 0.001, 'verbose': False}	Kfold(n_splits=5, random_state=42, shuffle=True)	0.642	-0.705	0.0059
68	Base completa; aumentando kfold de 5 para 10 - base de treinameto passa de 75% para 90%	38767	sklearn.linear_model.bayes.BayesianRidge	{'alpha_1': 1e-06, 'alpha_2': 1e-06, 'compute_score': False, 'copy_X': True, 'fit_intercept': True, 'lambda_1': 1e-06, 'lambda_2': 1e-06, 'n_iter': 300, 'normalize': False, 'tol': 0.001, 'verbose': False}	Kfold(n_splits=10, random_state=42, shuffle=True)	5.658	0.594	0.596
73	Base completa; shufflekfold 5 (teste=15%)	38767	sklearn.linear_model.ridge.Ridge	{'alpha': 0.5, 'copy_X': True, 'fit_intercept': True, 'max_iter': None, 'normalize': False, 'random_state': None, 'solver': 'auto', 'tol': 0.001}	ShuffleSplit(n_splits=5, random_state=42, test_size=0.15,train_size=None)	1.174	0.5941	0.595

Fonte: Elaborada pelo autor (2019).

APÊNDICE B – Detalhes da base espelho do projeto Cladop

O modelo relacional da Figura 13 descreve os dados que compõem a base espelho do projeto Cladop.

Figura 13 - Modelo relacional da base espelho com dados do sistema eTCE



Fonte: Elaborada pelo autor (2019).

Segue descrição detalhada do modelo.

Tabela tipo_documento_etce

Armazena os tipos associados ao Documento. Seus atributos são:

- Cod_tipo_documento_etce: Código do tipo de documento associado. Fonte documento_comprobatorio_dano.cod_tipo;
- Tipo_documento_etce: Descrição do tipo de documento associado. Fonte documento_comprobatorio_dano.cod_tipo;
- Se_exige_proporcao_min_texto: Identifica se o tipo exige que o documento tenha uma proporção mínima de páginas com texto em relação a imagens. Fonte:

tipo_documento_comprobatorio.se_exige_proporcao_min_texto. Constatado que o sistema não usa essa informação;

- Se_ativo: Indica se tipo está ativo. Fonte tipo_documento_comprobatorio.se_ativo;

Tabela dano

Armazena os danos ao Erário Público. Seus atributos são:

- Cod_dano: Código do Dano associado ao documento. Fonte: cod_dano.;
- Ano: Ano do dano. Fonte: ano;
- Num: Número do dano daquele ano - para controle do usuário. Fonte: num;
- Data_inicio_prazo_instauracao: Data de início da contagem do prazo de instauração. Formato: yyyymmdd. Fonte: data_inicio_prazo_instauracao;
- Data_instauracao: Data de início da instauração do dano. Formato: yyyymmdd. Fonte: data_instauracao;
- Data_atualizacao_debito: Data de atualização do valor do débito; Formato: yyyymmdd Fonte: data_atualizacao_debito;
- Ind_iniciativa_instauracao: Um dos valores da lista de valores possíveis de iniciativa de instauração. Fonte: ind_iniciativa_instauraca;
- Ind_motivo_instauracao: Um dos valores da lista de valores possíveis de motivos de instauração. Fonte: ind_motivo_instauracao;
- Ind_situacao: Situação do dano. Fonte: ind_situacao;
- Ind_submotivo_instauracao: Identifica o detalhamento do motivo de instauração selecionado. Fonte: ind_submotivo_instauracao;
- Qtd_doc_comprob_ausentes: Indica a quantidade de documentos comprobatórios ausentes. Fórmula: (case when dn.docs_comprobatorios_ausentes is null then 0 else (regexp_count(dn.docs_comprobatorios_ausentes, \;\) + 1) end) ;
- Valor_debito_atualiz_com_juros: Valor do débito atualizado com juros de mora. Fonte: valor_debito_atualiz_com_juros;
- Ind_origem_recursos: Um dos valores da lista de valores possíveis de origem de recurso. Fonte: ind_origem_recursos;
- Cod_unidade_instauradora: Indica unidade instauradora do dano. Fonte: cod_unidade_instauradora;
- Sigla_unidade_instauradora: Indica sigla de uma unidade instauradora do dano. Fonte: orgao_entidade.sigla para (tce. cod_orgao_externo);

Tabela documento_pdf_associado_dano

Armazena os documentos associados aos danos. Seus atributos são:

- Cod_documento: Código do documento associado a danos cadastrados no sistema eTCE. Fonte (banco de dados, owner TCE, formato nome da tabela.nome da coluna): documento_tramitavel.cod (para tce.documento_comprobatorio.cod_documento);
- Data_criacao: Data da criação da versão do arquivo eletrônico. Fonte: tcu.versao_arquivo_eletronico.data_criacao;
- Num_ordem: Ordem do documento de upload no Dano. Fonte: documento_comprobatorio_dano.ordem;
- Descricao_se_tipo_outros: Descrição do documento caso o tipo seja outros. Fonte: documento_comprobatorio_dano.descr_outros;
- Nome_arquivo: Nome do arquivo original, quando do upload para o sistema. Fonte: documento_comprobatorio_dano.nome_arquivo;
- Qtd_pagina_pdf: Qtd de página indicada pelo PDF. Valor atualizado durante pré-processamento do texto.
- Tamanho_arquivo: Indica o tamanho do arquivo em bytes. Fonte: tcu.versao_arquivo_eletronico.valor_tamanho_arquivo;

Tabela Data

Objetiva viabilizar consultas por partes de datas para as tabelas filhas. Armazena, quando sua criação, as datas de 1/1/1900 a 31/12/2100. Seus atributos são:

- Data: Data no formato: yyyymmdd
- Ano: Ano da data. Formato yyyy
- Mes: Mês da data. De 1 a 12.
- Dia_mes: Dia do mês da data. De 1 a 31.
- Dia_semana: Dia da semana da data. De 1 a 7.

Tabela metodo_extracao_texto

Armazena as diferentes formas de pré-processamento de texto usadas no projeto. Para cada método experimentado houve o registro do *texto processado* correspondente[58]. Durante a

[58] O texto filtrado e os atributos de qualidade do documento (*qtd_palavra_valida* por exemplo) que também dependem do método são armazenados na tabela *avaliacao_documento_pdf*.

evolução do projeto, métodos novos foram substituindo os anteriores. A forma de extração das palavras válidas também é um parâmetro para o treinamento denominado metodo_extracao_txt. Os seis métodos experimentados constam da Tabela 11 e retratam diferentes combinações das seguintes ações:

- Derivação de 7 classes substitutas de palavras (cpf/cnpj, números, datas, nomes de pessoas físicas, nomes e siglas de estados). Exemplificando, se no texto há o nome *rio de janeiro*, faz-se a substituição dessas 3 palavras por uma única palavra também considerada válida *classenomeuf*. O método atual (com identificação 7) usa todas as classes.

- Validação se um *token* é uma palavra correta se a mesma existir em um conjunto de palavras válidas. Conjuntos diferentes foram experimentados: palavras de acórdãos do TCU, palavras selecionadas da *wiki* em português e palavras selecionadas da *wiki* com acréscimo de abreviações e siglas típicas do negócio. O método atual usa o último conjunto, mais adequado ao negócio.

- Obtenção do texto do PDF, seja por OCR original ou por OCR extra com *tesseract*. Durante a evolução do projeto, dada a necessidade de um curto tempo de resposta do classificador para retornar uma previsão de tipo, abandonou-se a execução de OCR extra com *tesseract*, que demorava alguns minutos por documento.

A Tabela 14 apresenta detalhes dos métodos usados.

Tabela 14 - Métodos de pré-processamento usados no projeto

Código	Descrição
2	Conteúdo novo OCR com Tesseract 4.0; dicionário com palavras de acórdãos do TCU; 2 classes número e cpf/cnpj;
3	Conteúdo original com Pdftotext; dicionário com palavras de acórdãos do TCU; 2 classes número e cpf/cnpj;
4	Conteúdo novo OCR com Tesseract 4.0; dicionário com palavras selecionadas wiki; 2 classes número e cpf/cnpj;
5	Conteúdo OCR original com Pdftotext; dicionário com palavras selecionadas wiki; 2 classes número e cpf/cnpj;
6	Conteúdo OCR original com Pdftotext; dicionário com palavras selecionadas wiki após filtro; 2 classes número e cpf/cnpj;
7	Conteúdo OCR original com Pdftotext; dicionário com palavras selecionadas wiki e termos comuns do negócio; 7 classes;

Fonte: Elaborada pelo autor (2019).

Tabela Avaliacao_documento_pdf

Armazena indicadores de qualidade dos textos dos documentos apurados com a aplicação do método indicado. Seus atributos são:

- Cod_metodo_extracao_texto_pdf: Código do método de pré-processamento usado.
- Cod_documento: Código do documento avaliado pelo método;

- Data_apuracao: Data da realização da avaliação;
- Texto: Texto resultante do pré-processamento;
- Qtd_pagina_pdf: Indica a quantidade de página indicada pelo PDF. Redundância da tabela documento_pdf_dano para otimizar cálculo de métricas;
- Tamanho_arquivo: Indica o tamanho do arquivo em bytes, obtido da tabela do GED TCU.VERSAO_ARQUIVO_ELETRONICO, coluna valor_tamanho_arquivo. Redundância da tabela documento_pdf_dano para otimizar cálculo de métricas;
- Qtd_char_vogal: Indica a quantidade de caracteres que são vogais válidas (acentuadas ou não) encontrados no texto gerado pelo método indicado. Contraexemplo: "è";
- Qtd_char_consoante: Indica a quantidade de caracteres que são consoantes válidas encontrados no texto gerado pelo método indicado. Contraexemplo: "ñ";
- Qtd_char_letra: Indica a quantidade de caracteres que são vogais ou consoantes.
- Qtd_token: Indica a quantidade de tokens (pedaços de texto) identificados no texto gerado pelo método indicado. Considera como pedaço de texto toda sequência de caracteres (tamanho 1 para cima) separada por \s (espaço em branco /n, tabulação /t, etc). Exemplo, na frase "ada fdç/as ADA sem amor" tem 5 pedaços. Em princípio, é o somatório de qtd_palavra + qtd_numero + qtd_token_di';
- Qtd_token_palavra: Indica a quantidade de ocorrências (inclui repetições) de palavras (sequência de letras) válidas ou não;
- Qtd_token_palavra_valida: Indica a quantidade de ocorrências (inclui repetições) de palavras validas (confirmadas junto a uma fonte confiável – palavras de acórdãos do TCU) identificadas no texto gerado pelo método indicado. Exemplo: considera palavra "exceção" e "excecao". Pressupõe pré-processamento que transforma as palavras em lower case;
- Qtd_token_dif: Indica a quantidade de tokens diferentes, que não são número nem palavra no texto gerado pelo método indicado. Exemplo, na frase "ada fdç/as ADA sem amor" tem um token: “fdç/a”;
- Qtd_token_numero: Indica a quantidade de números encontrados no texto gerado pelo método indicado. Considera número um agregado de dígitos (podendo haver ".", "-", "/" e "," entre eles). A frase "Hoje 33 a exceção 23 tem sim R$212,01 valor 629,456-23 de 1998", por exemplo, tem 5 números;
- Qtd_token_num_cpf_cnpj: Indica a quantidade de CPFs e CNPJs no texto gerado pelo método indicado;

- Qtd_char: Indica a quantidade total de caracteres encontrados no texto gerado pelo método indicado. Conta brancos repetidos (pressupõe que não haja pré-processamento que unifique brancos concatenados em um apenas);
- Qtd_char_numero: Indica a quantidade de caracteres que são números encontrados no texto gerado pelo método indicado;
- Qtd_char_espaco: Indica a quantidade de caracteres de espaço (\s: branco, \n, etc) encontrados no texto gerado pelo método indicado. Conta repetições seguidas;
- Qtd_char_pontuacao: Quantidade de caracteres de pontuação encontrados no texto gerado pelo método indicado. Considera caractere pontuação ".,?!". Pressupõe pré-processamento com lowercase. Na frase: "Hoje a exceção ~! tem sim% valor" há 3 (~!%);
- Qtd_char_dif: Indica a quantidade de caracteres especiais encontrados no texto gerado pelo método indicado. Considera caractere dif (diferente) o que não é letra válida, número, pontuação e espaço';
- Qtd_trio_consoante: Indica a quantidade de trios de consoante. Exemplo: mns, cdG;
- Qtd_trio_char_repetido_nesp: Indica a quantidade de trios de caracteres repetidos encontrados. Desconsidera espaço. Exemplo: aaa, eee, Ccc, ddd;
- Qtd_char_separador: Indica a quantidade de caracteres separadores (ex.: /[{) que são números encontrados no texto gerado pelo método indicado;
- Qtd_char_simbolo: Indica a quantidade de caracteres que não estão nas outras categorias, mas são símbolos conhecidos (ex.: +*#$) que são números encontrados no texto gerado pelo método indicado;
- Qtd_linha: Indica a quantidade de linhas no texto gerado pelo método indicado;
- Qtd_data: Indica a quantidade de data no texto (passou a gravar após 8/7/2019);
- Qtd_parte_nome: Indica a quantidade de partes de nomes (nome ou sobrenomes mais comuns) no texto (passou a gravar após 8/7/2019);
- Qtd_sigla_uf: Indica a quantidade de sigla uf no texto (passou a gravar após 8/7/2019);
- Qtd_nome_uf: Indica a quantidade de nome uf no texto (passou a gravar após 8/7/2019);
- Percent_token_dif: Indica o percentual de tokens diferente em relação ao total tokens token (qtd_token_dif/qtd_token). Redundância para otimizar o cálculo;
- Tamanho_token_media: Indica o tamanho médio dos tokens identificado no texto gerado pelo método indicado;
- Tamanho_token_std: Indica o desvio padrão do tamanho dos tokens identificado no texto gerado pelo método indicado;

- Prop_letra_palavra_valida: Indica a quantidade de letras do documento em relação à quantidade de palavras válidas;
- Percent_palavra_valida_token: Percentual de tokens que são palavras válidas;
- Percent_numero_token: Percentual de tokens que são palavras números;
- Percent_outros_token: Percentual de tokens que não são palavras válidas nem números;
- Percent_espaco_char: Proporção de espaços entre os caracteres;
- Percent_token_dif: Indica o percentual de tokens diferentes em relação ao total de tokens (qtd_token_dif/qtd_token). Redundância para otimizar o cálculo;
- Percent_char_dif: Indica o percentual de char diferente em relação ao total de char. (qtd_char_dif/qtd_char). Redundância para otimizar o cálculo;
- Percent_char_seq_invalida: Indica o percentual de char diferente de espaço que compõe sequências avaliadas como inválidas em relação ao total de char. ((qtd_trio_consoante + qtd + trio_char_repetido_nesp + qtd_trio_outro) * 3 / (qtd_char - qtd_char_espaco)). Redundância para otimizar o cálculo;
- Qtd_xxx_ppag: Todos os indicadores de quantidade (XXX, ver acima) divididos pelo número de páginas do documento.
- Qtd_xxx_plin: Todos os indicadores de quantidade (XXX, ver acima) divididos pelo número de linhas do documento.

APÊNDICE C – Fluxo de processamento da rotina proposta para monitoramento de desempenho do modelo de classificação multi-classe

Objetivo

Monitorar o desempenho pela métrica de acurácia (micro) do modelo Cladop em produção e, se necessário, treinar nova subversão do modelo.

A rotina pode ser aplicada a outros projetos de classificação que utilizem a infraestrutura de rastro do Cladop (modelo com metadados dos projetos e dados de treinamentos associados às versões em produção)[59].

A ideia de se construir uma rotina genérica objetiva demonstrar que a estrutura de rastro criada para o registro de treinamento do modelo, além de ser uma rica documentação do projeto, pode ser aplicada no monitoramento do modelo (fase Implantação do CRISP-DM). Demonstra que o rastro pode integrar a fase de desenvolvimento a fase de pós-desenvolvimento (monitoramento).

Detalhes da rotina e o código encontram-se publicados em https://gitlab.com/MarcusBorela/rastro-dm.git, na pasta Monitoramento.

Detalhes de execução

Periodicidade mensal: sempre no dia primeiro do mês, às 2 horas (comando *cron* do *linux*).

Máquina: 10.22.9.36 (Estação número 1 de desenvolvimento de serviços cognitivos do Seint – com GPU)

Os parâmetros para o monitoramento do projeto são obtidos de tabelas de metadados armazenadas no banco *oracle* (produção).

Em caso de erro de processamento, um e-mail de erro de processamento é enviado para o e-mail da equipe de projeto (que consta nos metadados do projeto) e é registrado o erro na no campo referente ao último registro de monitoramento.

[59] A versão atual da rotina (*Cladop_monitoramento.ipynb*) só processa o monitoramento do primeiro projeto retornado, no caso o Cladop (código = 1). Para se aplicar a outros projetos, faz-se necessário implementar o tratamento de múltiplos projetos com repetição dos passos por projeto.

São duas tabelas atributivas do projeto que estão diretamente ligadas à rotina de monitoramento. Seguem abaixo os comentários das tabelas e de suas colunas.

Para simplificação, não consta da relação a coluna *cod_projeto* que é a chave primária das tabelas (é previsto um registro apenas por projeto, por isso a denominação *tabela atributiva*, uma extensão de atributos que poderiam estar na tabela projeto_classificador, mas, por opção de modelagem, fez-se a separação por clareza e por facilidade de evolução futura).

A Tabela 15 apresenta a estrutura da tabela *monitoramento_mensal*, que contém os parâmetros que orientam como se dá o monitoramento mensal do projeto. Contém também informações da última execução.

Tabela 15 - Estrutura da tabela monitoramento_mensal

Coluna	Comentário
num_dias_periodo_apuracao	Indica o número de dias a considerar na apuração da acurácia efetiva do modelo em produção. Por exemplo, 30 dias.
qtd_min_registro_novo	Indica a quantidade mínima de registros novos (inseridos após o modelo atualmente em produção) necessária para se efetuar novo treinamento. Parâmetro que apoia a definição da viabilidade de um novo treinamento.
qtd_min_registro_treinamento	Indica a quantidade de registros necessária para o novo treinamento. Parâmetro usado na seleção de dados para o novo treinamento. se for necessário.
qtd_min_registro_por_classe	Indica a quantidade mínima de registros por classe. Parâmetro usado na seleção de dados para o novo treinamento, se for necessário.
probab_acc_superior_limite_min	Indica a probabilidade desejada de que a acurácia apurada esteja acima do limite mínimo de acurácia a ser estipulado. Parâmetro que apoia a definição dos limites de acurácia aceitáveis.
pct_populacao_acc_sup_lim_min	Indica o percentual da população esperado que tem acurácia superior ao mínimo a ser estipulado. Parâmetro que apoia a definicão dos limites de acurácia aceitáveis para o período.
descr_comando_sel_acc_periodo	Critério no formato *sql* usado para seleção da acurácia no período monitorado. Necessita ter no texto o parâmetro num_dias_periodo_apuracao e retornar: qtd_registro_periodo e qtd_acerto_tipo_periodo. Se houver registros não classificáveis (tipo "outros" ou outra situação), pode trazer também: qtd_registro_nao_classificavel. Nesse caso, qtd_registro_classificavel será qtd_registro_periodo - qtd_registro_nao_classificavel
descr_comando_sel_data	Critério no formato sql usado para definir a data limite de seleção de dados mais recentes para novo treinamento. Deve retornar a maior data que tenha registros em número suficiente para o treinamento, ou seja, que atenda aos parâmetros passados: qtd_min_registro_treinamento e qtd_min_registro_por_classe. Esses parâmetros, se necessários para o projeto, devem estar no sql antecedidos por ":". Deve retornar data_limite em formato date.
descr_comando_sel_qtd_reg_novo	Critério no formato sql usado para seleção da quantidade de documentos novos que ainda não participaram do treinamento do modelo em produção. Deve retornar coluna qtd_registro.
descr_codigo_processa_reg_novo	Código que implementa o processamento de registros novos (se houver necessidade para o classificador). Regra de formação: cada linha precisa ser um comando *python*.
dthora_ultima_execucao	Indica a data e hora da última execução do monitoramento.
descr_situacao_ultima_execucao	Texto indicativo da situação, do resultado da última execução.

Fonte: Elaborada pelo autor (2019).

A Tabela 16 traz a estrutura da tabela *forma_apuracao_acuracia_teste*, que indica a forma como se dá a apuração da acurácia em dados de testes do modelo. Somente será considerada a apuração por validação cruzada (*RepeatedKFold*). Os valores das colunas quando multiplicados indicam o número de amostras usadas na apuração das métricas. Esse número é usado pela rotina de monitoramento para a determinação dos limites mínimos aceitáveis.

Tabela 16 - Estrutura da tabela forma_apuracao_acuracia_teste

Coluna	Comentário
num_particao	Indica o número de partições (folds) a ser usado no método de validação cruzada Repeated*Kfold* (*python*).
num_repeticao	Indica o número de repetições a ser usado no método de validação cruzada Repeated*Kfold* (*python*).

Fonte: Elaborada pelo autor (2019).

Fluxo de Processamento

A Tabela 17 abaixo indica os passos do fluxo de processamento da rotina de monitoramento.

Tabela 17 - Fluxo de processamento da rotina de monitoramento integrada ao rastro

Passo	Parâmetros de entrada e/ou condição	Atividade	Dados gerados
1		Obter dados modelo em produção e parâmetros de monitoramento do projeto (conecta com o banco e busca metadados do modelo)	* No dicionário parametro_monitoramento: # dados básicos do projeto cod_projeto sigla_projeto descr_projeto email_equipe_modelo email_equipe_sistema_uso # dados da versão em produção num_versao num_subversao percent_acuracia_teste percent_acuracia_teste_std num_amostra_apuracao # dados de monitoramento ##parâmetros de apuração de acurácia num_dias_periodo_apuracao probab_acc_superior_limite_min pct_populacao_acc_sup_lim_min descr_comando_sel_acc_periodo ##critérios para novo treinamento de modelo descr_comando_sel_data qtd_min_registro_treinamento qtd_min_registro_por_classe qtd_min_registro_novo descr_comando_sel_qtd_reg_novo # forma de apuração de acurácia em novo treinamento nova_apuracao_num_particao nova_apuracao_num_repeticao
2	num_amostra_apuracao probab_acc_superior_limite_min pct_populacao_acc_sup_lim_min	Definir limites aceitáveis para aceitação da acurácia apurada no período (considera a Tabela A-7 do livro Estatística Experimental - Natrella, M. G., Estatística Experimental, NBS Handbook 91, 1963)	* parametro_monitoramento: acc_limite_minimo_aceitavel acc_limite_maximo_esperado

		A versão atual da rotina trabalha com probab_acc_superior_limite_min = 99% e pct_populacao_acc_sup_lim_min = 99,9%	num_amostra_apuracao_estimada (se não for possível encontrar os limites na tabela, assume-se o tamanho de amostra imediatamente inferior ao real que constar da tabela)
3	num_dias_periodo_apuracao descr_comando_sel_acc_periodo	Apurar acurácia no período	* dicionário apuracao_periodo: qtd_registro qtd_acerto_tipo percent_acuracia_apurada qtd_registro_classificavel * Se houver registros não classificáveis: qtd_registro_nao_classificavel percent_nao_classificavel
3.1	Se percent_acuracia_apurada_periodo >= acc_limite_minimo_aceitavel	*Enviar e-mail informando: "O modelo está com acurácia dentro do esperado" ou, conforme o caso, "O modelo está com acurácia dentro do esperado e superior ao limite máximo esperado"	
4		Processar novos documentos Documentos novos que não foram usados no treinamento do modelo em produção. Trata-se do pré-processamento dos textos dos novos registros para treinamento do modelo. No caso do Cladop, atualiza-se uma base espelho com dados base para o treinamento (que espelha o sistema em produção) e faz-se o pré-processamento do texto dos documentos.	
5	descr_comando_sel_qtd_reg_novo	Apurar quantitativo de registros não usados no treinamento do modelo em produção No caso do Cladop, contabilizam-se os documentos inseridos após a data de geração do modelo atual	* apuracao_periodo: qtd_registro_novo
5.1	Se qtd_registro_novo <= qtd_min_registro_novo	*Enviar e-mail informando "Embora a acurácia tenha sido inferior ao esperado, não foram encontrados registros não usados no treinamento do modelo em produção em número suficiente para o treinamento. Se necessário, os parâmetros do projeto podem ser revistos!"	
6	qtd_min_registro_treinamento qtd_min_registro_por_classe descr_comando_sel_data	Definir data limite para seleção de dados para novo treinamento (encontra a maior data que tenha registros em número suficiente para o treinamento, ou seja, que atenda aos parâmetros passados)	* apuracao_periodo: data_inicio_selecao_treinamento data_inicio_selecao_treinamento_formato_oracle (que pode ser colocado em comando sql)
6.1	Se não encontrada data_inicio_selecao_treinamento	*Enviar e-mail informando:	

		"Não foi possível identificar uma data de corte para seleção de registros para um novo treinamento que atenda aos requisitos cadastrados para o projeto. Se necessário, os parâmetros do projeto podem ser revistos!"	
7	data_inicio_selecao_treinamento_formato_oracle nova_apuracao_num_particao nova_apuracao_num_repeticao cod_treinamento_geracao_versao	Apurar acurácia do modelo com nova seleção de dados (treinar novo modelo) Treina o modelo com validação cruzada (Repeatd*Kfold*) usando o número de partições e de repetições especificado para o projeto.	* novo_modelo: percent_acuracia percent_acuracia_std cod_treino_teste
7.1	Se novo_modelo.percent_acuracia <= percent_acuracia_teste ou (novo_modelo.percent_acuracia - novo_modelo.percent_acuracia _std) <= (percent_acuracia_teste - percent_acuracia_teste_std)	*Enviar e-mail informando: 'O modelo treinado não alcançou acurácia superior ao modelo em produção e foi descartado. Sugere-se avaliar a revisão dos parâmetros de treinamento do modelo atual.'	
8	novo_modelo.cod_treino_teste	Gerar nova subversão do modelo (treino final sem separação de dados de teste)	* novo_modelo: percent_acuracia_validacao ** cod_treino_geracao
9	cod_projeto num_versao num_subversao cod_treino_teste cod_treino_geracao	Registrar nova subversão nos metadados do projeto	Subversão registrada nos metadados do projeto
9.1		*Enviar e-mail informando: 'Novo modelo alcançou acurácia superior ao modelo atual foi treinado e cadastrado como nova subversão nos metadados de projetos. Sugere-se, após os devidos testes, a sua implantação."	

Fonte: Elaborada pelo autor (2019).

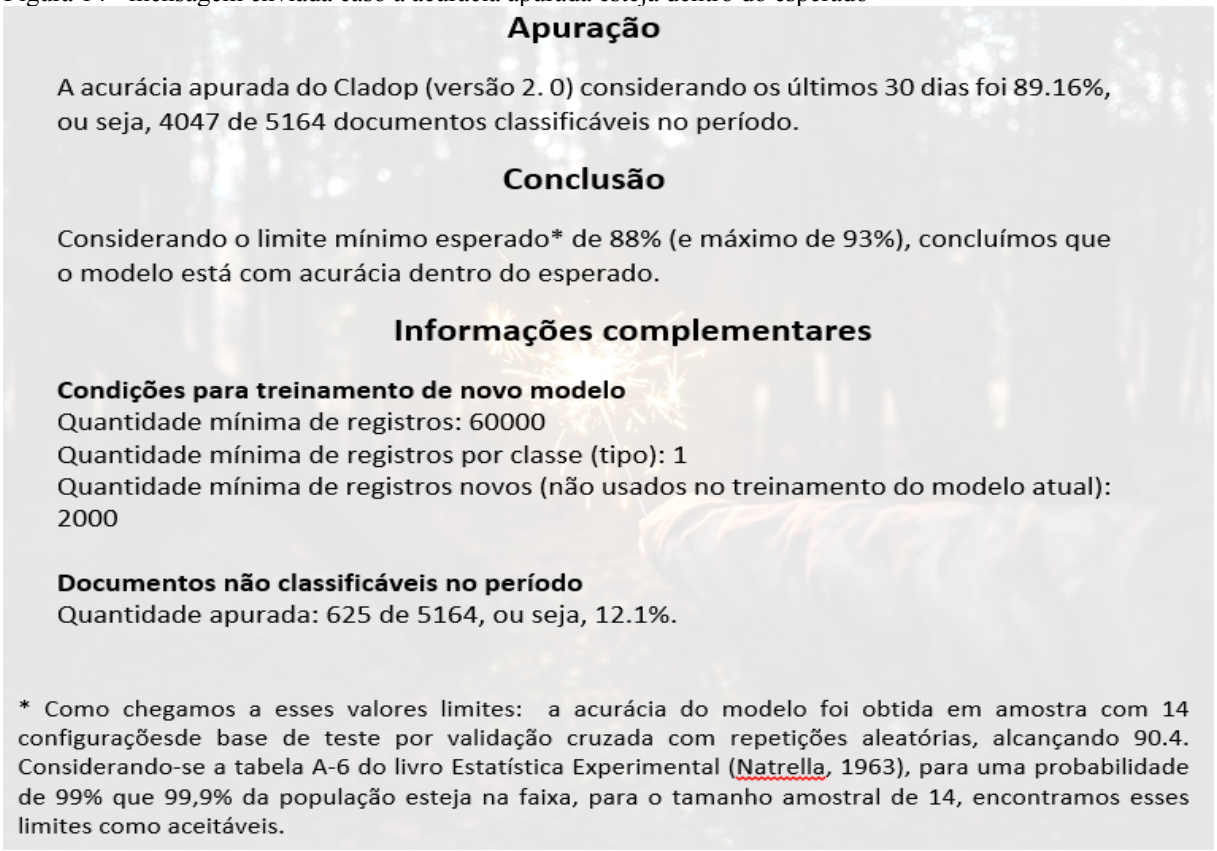
* Os e-mails são enviados para e-mail indicado no metadado do projeto: email_equipe_modelo[60]. São dois *templates* que devem ser usados conforme tenha ocorrido ou não treinamento de um novo modelo: "monitoramento com treinamento.docx" ou "monitoramento sem treinamento.docx". Adicionalmente, registra-se o resultado do monitoramento e a data e a hora nos campos reservados para a última execução do monitoramento (colunas dthora_ultima_execucao e descr_situacao_ultima_execucao da tabela monitoramento_mensal) e encerra-se o fluxo de processamento.

** retornado esse parâmetro se a técnica empregada for de rede neural, caso do Cladop. Para algoritmos *shallow*, não há necessidade de se estipular um conjunto de validação, podendo ser gerada uma versão que engloba todos os dados disponíveis.

Exemplos de e-mails enviados automaticamente[61]

i) Caso a acurácia apurada esteja dentro do esperado:

Figura 14 - mensagem enviada caso a acurácia apurada esteja dentro do esperado



Apuração

A acurácia apurada do Cladop (versão 2. 0) considerando os últimos 30 dias foi 89.16%, ou seja, 4047 de 5164 documentos classificáveis no período.

Conclusão

Considerando o limite mínimo esperado* de 88% (e máximo de 93%), concluímos que o modelo está com acurácia dentro do esperado.

Informações complementares

Condições para treinamento de novo modelo
Quantidade mínima de registros: 60000
Quantidade mínima de registros por classe (tipo): 1
Quantidade mínima de registros novos (não usados no treinamento do modelo atual): 2000

Documentos não classificáveis no período
Quantidade apurada: 625 de 5164, ou seja, 12.1%.

* Como chegamos a esses valores limites: a acurácia do modelo foi obtida em amostra com 14 configuraçõesde base de teste por validação cruzada com repetições aleatórias, alcançando 90.4. Considerando-se a tabela A-6 do livro Estatística Experimental (Natrella, 1963), para uma probabilidade de 99% que 99,9% da população esteja na faixa, para o tamanho amostral de 14, encontramos esses limites como aceitáveis.

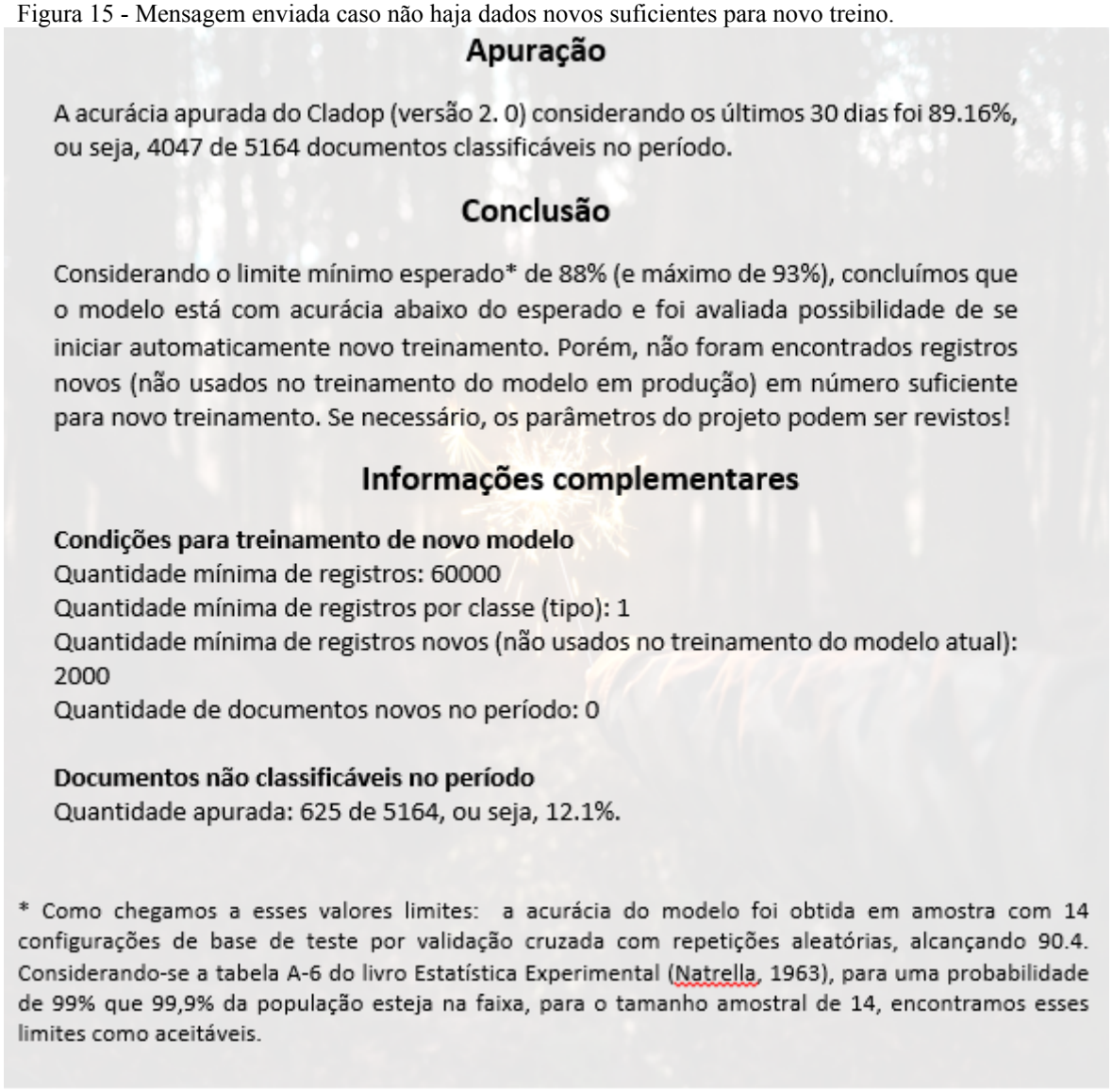
Fonte: Elaborada pelo autor (2019).

[60] Versão futura também enviará términos com sucesso para a equipe que cuida do sistema que usa o classificador (coluna *email_equipe_sistema_uso da tabela projeto_classificador*).

[61] Gerados durante os testes da rotina.

ii) Caso a acurácia apurada não esteja dentro do esperado, mas não haja dados novos em número suficiente para o treinamento:

Figura 15 - Mensagem enviada caso não haja dados novos suficientes para novo treino.



Apuração

A acurácia apurada do Cladop (versão 2. 0) considerando os últimos 30 dias foi 89.16%, ou seja, 4047 de 5164 documentos classificáveis no período.

Conclusão

Considerando o limite mínimo esperado* de 88% (e máximo de 93%), concluímos que o modelo está com acurácia abaixo do esperado e foi avaliada possibilidade de se iniciar automaticamente novo treinamento. Porém, não foram encontrados registros novos (não usados no treinamento do modelo em produção) em número suficiente para novo treinamento. Se necessário, os parâmetros do projeto podem ser revistos!

Informações complementares

Condições para treinamento de novo modelo
Quantidade mínima de registros: 60000
Quantidade mínima de registros por classe (tipo): 1
Quantidade mínima de registros novos (não usados no treinamento do modelo atual): 2000
Quantidade de documentos novos no período: 0

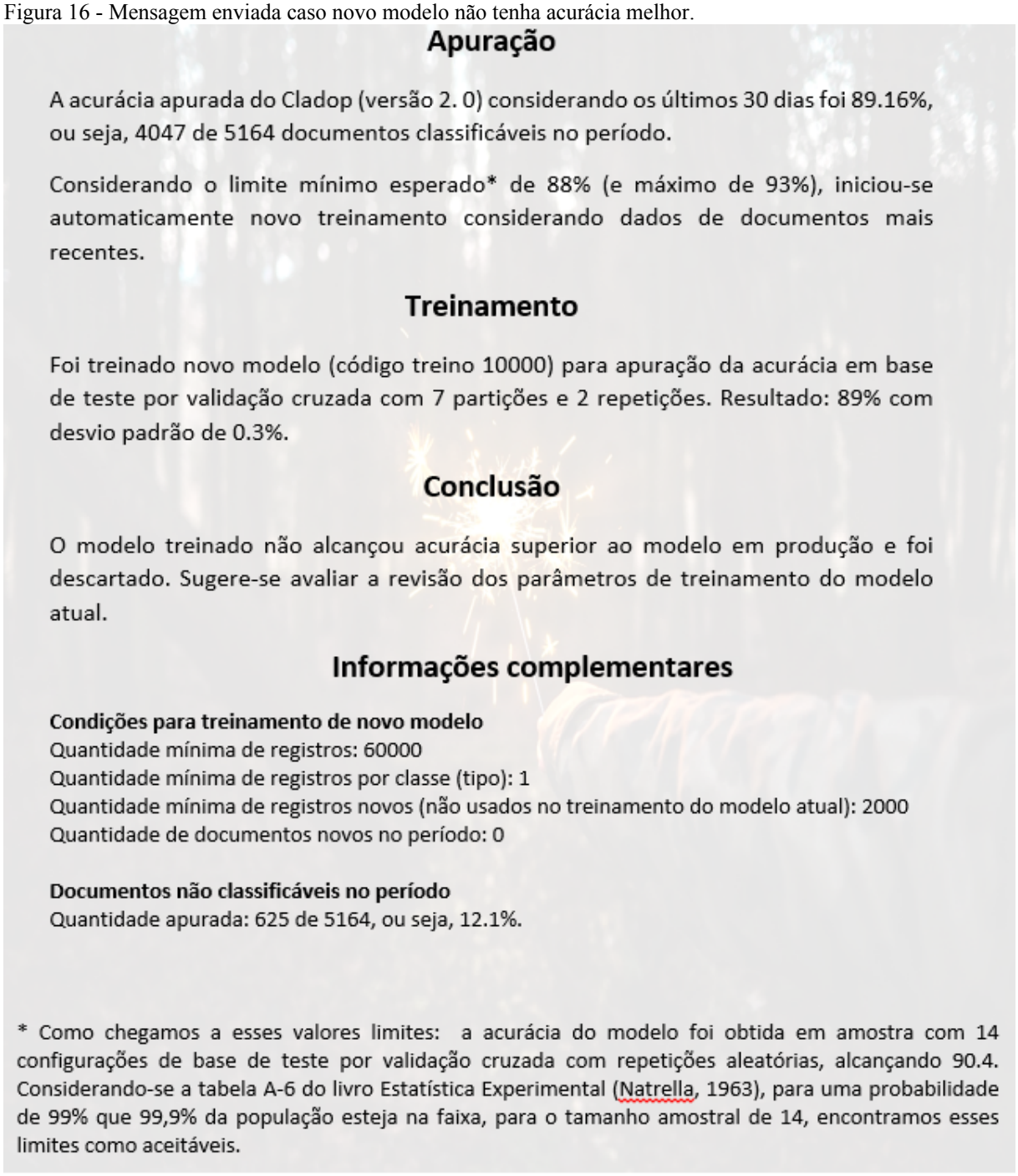
Documentos não classificáveis no período
Quantidade apurada: 625 de 5164, ou seja, 12.1%.

* Como chegamos a esses valores limites: a acurácia do modelo foi obtida em amostra com 14 configurações de base de teste por validação cruzada com repetições aleatórias, alcançando 90.4. Considerando-se a tabela A-6 do livro Estatística Experimental (Natrella, 1963), para uma probabilidade de 99% que 99,9% da população esteja na faixa, para o tamanho amostral de 14, encontramos esses limites como aceitáveis.

Fonte: Elaborada pelo autor (2019).

iii) Caso a acurácia apurada não esteja dentro do esperado e novo modelo treinado não tenha acurácia superior ao modelo atual:

Figura 16 - Mensagem enviada caso novo modelo não tenha acurácia melhor.



Apuração

A acurácia apurada do Cladop (versão 2. 0) considerando os últimos 30 dias foi 89.16%, ou seja, 4047 de 5164 documentos classificáveis no período.

Considerando o limite mínimo esperado* de 88% (e máximo de 93%), iniciou-se automaticamente novo treinamento considerando dados de documentos mais recentes.

Treinamento

Foi treinado novo modelo (código treino 10000) para apuração da acurácia em base de teste por validação cruzada com 7 partições e 2 repetições. Resultado: 89% com desvio padrão de 0.3%.

Conclusão

O modelo treinado não alcançou acurácia superior ao modelo em produção e foi descartado. Sugere-se avaliar a revisão dos parâmetros de treinamento do modelo atual.

Informações complementares

Condições para treinamento de novo modelo
Quantidade mínima de registros: 60000
Quantidade mínima de registros por classe (tipo): 1
Quantidade mínima de registros novos (não usados no treinamento do modelo atual): 2000
Quantidade de documentos novos no período: 0

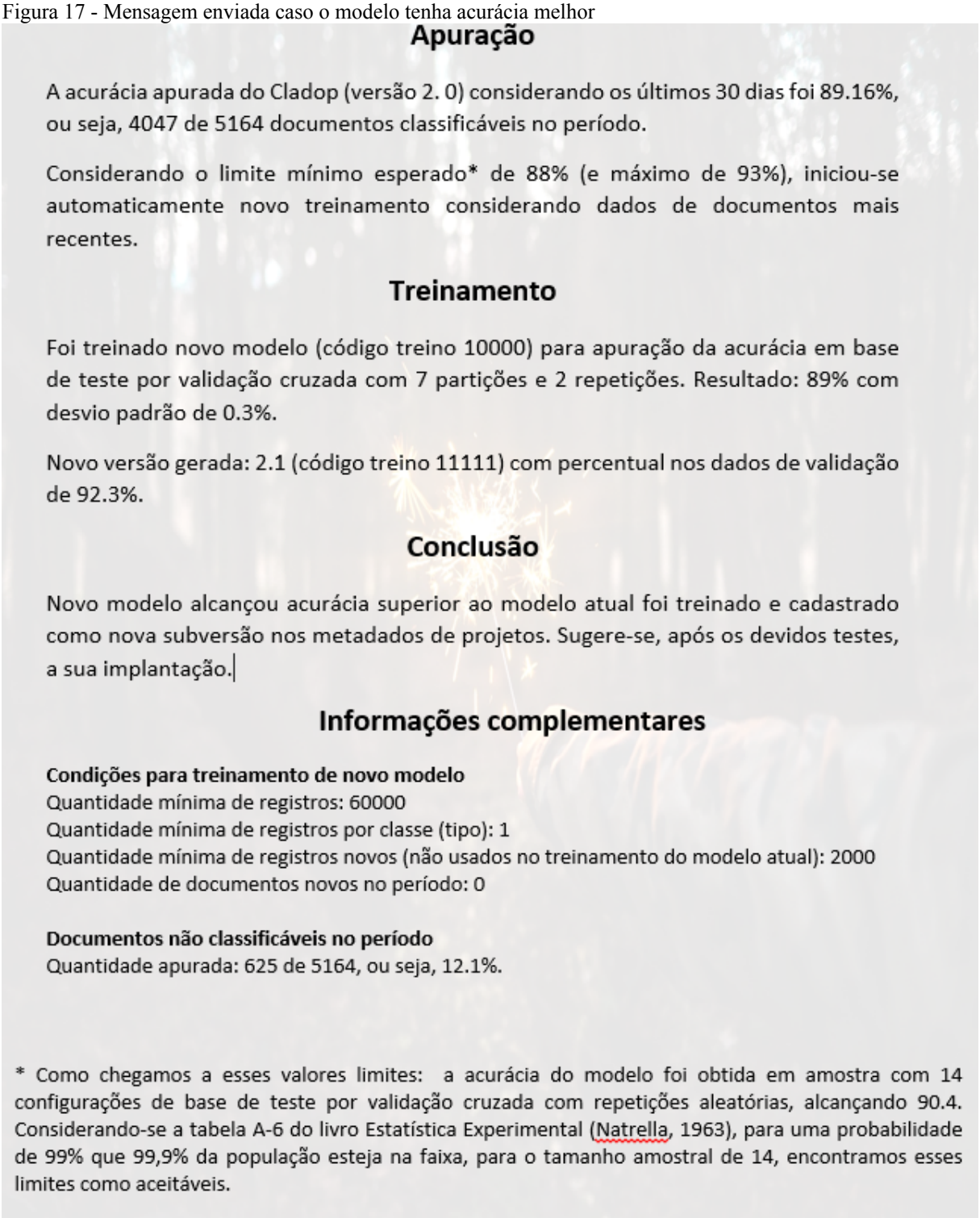
Documentos não classificáveis no período
Quantidade apurada: 625 de 5164, ou seja, 12.1%.

* Como chegamos a esses valores limites: a acurácia do modelo foi obtida em amostra com 14 configurações de base de teste por validação cruzada com repetições aleatórias, alcançando 90.4. Considerando-se a tabela A-6 do livro Estatística Experimental (Natrella, 1963), para uma probabilidade de 99% que 99,9% da população esteja na faixa, para o tamanho amostral de 14, encontramos esses limites como aceitáveis.

Fonte: Elaborada pelo autor (2019).

iv) Caso a acurácia apurada não esteja dentro do esperado e novo modelo treinado alcance acurácia superior ao modelo atual:

Figura 17 - Mensagem enviada caso o modelo tenha acurácia melhor



Apuração

A acurácia apurada do Cladop (versão 2. 0) considerando os últimos 30 dias foi 89.16%, ou seja, 4047 de 5164 documentos classificáveis no período.

Considerando o limite mínimo esperado* de 88% (e máximo de 93%), iniciou-se automaticamente novo treinamento considerando dados de documentos mais recentes.

Treinamento

Foi treinado novo modelo (código treino 10000) para apuração da acurácia em base de teste por validação cruzada com 7 partições e 2 repetições. Resultado: 89% com desvio padrão de 0.3%.

Novo versão gerada: 2.1 (código treino 11111) com percentual nos dados de validação de 92.3%.

Conclusão

Novo modelo alcançou acurácia superior ao modelo atual foi treinado e cadastrado como nova subversão nos metadados de projetos. Sugere-se, após os devidos testes, a sua implantação.

Informações complementares

Condições para treinamento de novo modelo
Quantidade mínima de registros: 60000
Quantidade mínima de registros por classe (tipo): 1
Quantidade mínima de registros novos (não usados no treinamento do modelo atual): 2000
Quantidade de documentos novos no período: 0

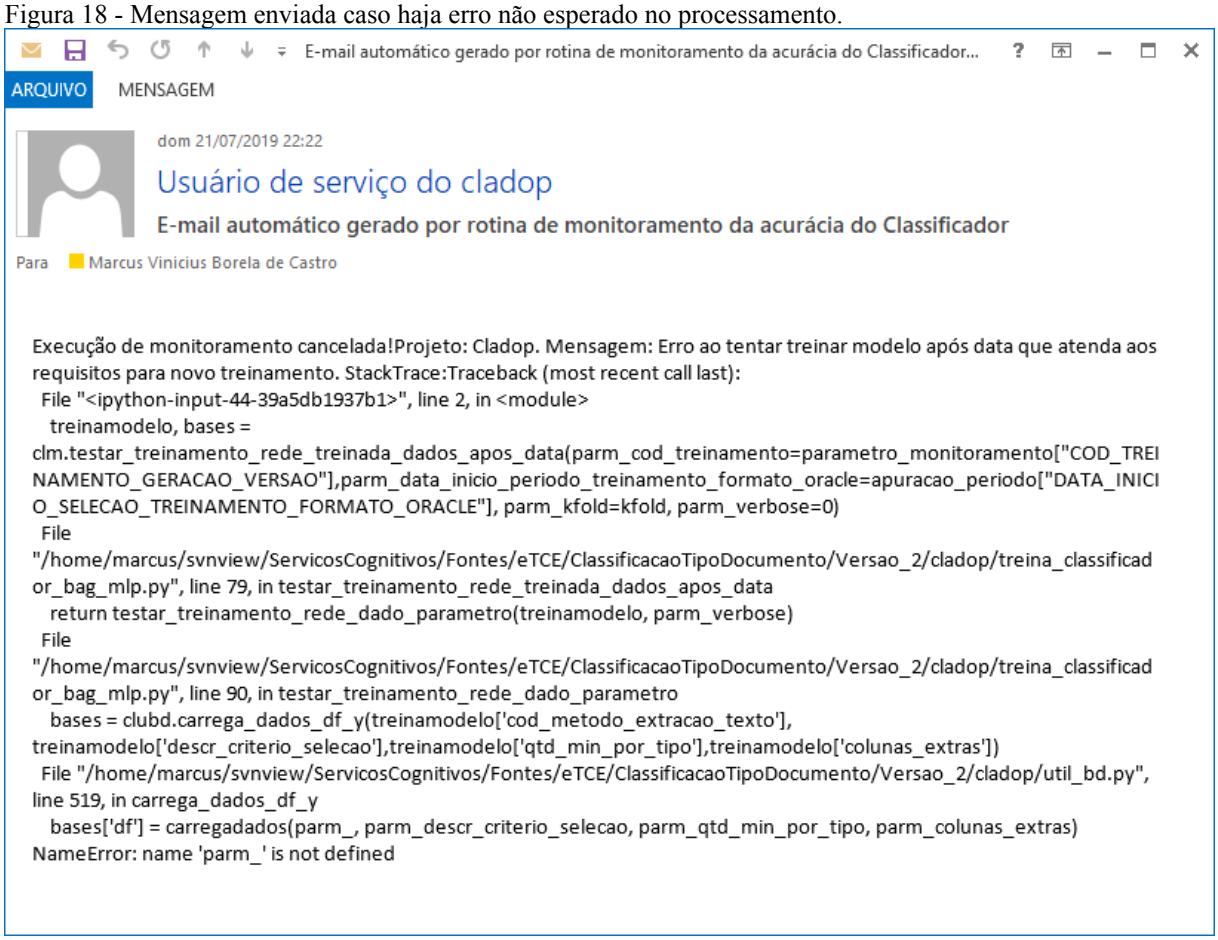
Documentos não classificáveis no período
Quantidade apurada: 625 de 5164, ou seja, 12.1%.

* Como chegamos a esses valores limites: a acurácia do modelo foi obtida em amostra com 14 configurações de base de teste por validação cruzada com repetições aleatórias, alcançando 90.4. Considerando-se a tabela A-6 do livro Estatística Experimental (Natrella, 1963), para uma probabilidade de 99% que 99,9% da população esteja na faixa, para o tamanho amostral de 14, encontramos esses limites como aceitáveis.

Fonte: Elaborada pelo autor (2019).

v) Caso haja erro na execução da rotina:

Figura 18 - Mensagem enviada caso haja erro não esperado no processamento.



Fonte: Elaborada pelo autor (2019).

Templates usados para envio de e-mails:

A imagem usada como fundo nos *templates* pode ser obtida em: https://images.app.goo.gl/LHPdZX3A5CvDofUj9 (Photo by Jamie Street on Unsplash-jamie-street-202592-unsplash)

i) Se não houver treinamento de modelo:

Apuração

A acurácia apurada do {{sigla_projeto}} (versão {{num_versao}}.{{num_subversao}}) considerando os últimos {{num_dias_periodo_apuracao}} dias foi {{percent_acuracia_apurada}}%, ou seja, {{qtd_acerto_tipo_periodo}} de {{qtd_registro_classificavel}} documentos classificáveis no período.

Conclusão

Considerando o limite mínimo esperado* de {{acc_limite_minimo_aceitavel}}% (e máximo de {{acc_limite_maximo_esperado}}%), concluímos que {{mensagem}}

Informações complementares

Condições para treinamento de novo modelo

Quantidade mínima de registros: {{qtd_min_registro_treinamento}}

Quantidade mínima de registros por classe (tipo): {{qtd_min_registro_por_classe}}

Quantidade mínima de registros novos (não usados no treinamento do modelo atual): {{qtd_min_registro_novo}} {% if qtd_registro_novo != '****' %}

Quantidade de documentos novos no período: {{qtd_registro_novo}} {% endif %}

{% if qtd_registro_nao_classificavel != '****' %}

Documentos não classificáveis no período

Quantidade apurada: {{qtd_registro_nao_classificavel}} de {{qtd_registro_periodo}}, ou seja, {{percent_nao_classificavel}}%.

{% endif %}

* Como chegamos a esses valores limites: a acurácia do modelo foi obtida em amostra com {{num_amostra_apuracao}} configurações de base de teste por validação cruzada com repetições aleatórias, alcançando {{percent_acuracia_teste}}. Considerando-se a tabela A-6 do livro Estatística Experimental (Natrella, 1963), para uma probabilidade de 99% que 99,9% da população esteja na faixa, para o tamanho amostral de {{num_amostra_apuracao_assumida_definicao_limite_acc}}, encontramos esses limites como aceitáveis.

ii) Se houver treinamento de modelo

Apuração

A acurácia apurada do {{sigla_projeto}} (versão {{num_versao}}.{{num_subversao}}) considerando os últimos {{num_dias_periodo_apuracao}} dias foi {{percent_acuracia_apurada}}%, ou seja, {{qtd_acerto_tipo_periodo}} de {{qtd_registro_classificavel}} documentos classificáveis no período.

Considerando o limite mínimo esperado* de {{acc_limite_minimo_aceitavel}}% (e máximo de {{acc_limite_maximo_esperado}}%), iniciou-se automaticamente novo treinamento considerando dados de documentos mais recentes.

Treinamento

Foi treinado novo modelo (código treino {{cod_treino_teste}}) para apuração da acurácia em base de teste por validação cruzada com {{nova_apuracao_num_particao}} partições e {{nova_apuracao_num_repeticao}} repetições. Resultado: {{percent_acuracia_teste_novo_modelo}}% com desvio padrão de {{percent_acuracia_teste_std_novo_modelo}}%. {% if cod_treino_geracao != '****' %}

Novo versão gerada: {{num_versao}}.{{num_subversao + 1}} (código treino {{cod_treino_geracao}}) com percentual nos dados de validação de {{percent_acuracia_validacao_novo_modelo}}%. {% endif %}

Conclusão

{{mensagem}}

Informações complementares

Condições para treinamento de novo modelo

Quantidade mínima de registros: {{qtd_min_registro_treinamento}}

Quantidade mínima de registros por classe (tipo): {{qtd_min_registro_por_classe}}

Quantidade mínima de registros novos (não usados no treinamento do modelo atual): {{qtd_min_registro_novo}} {% if qtd_registro_novo != '****' %}

Quantidade de documentos novos no período: {{qtd_registro_novo}} {% endif %}

{% if qtd_registro_nao_classificavel != '****' %}

Documentos não classificáveis no período

Quantidade apurada: {{qtd_registro_nao_classificavel}} de {{qtd_registro_periodo}}, ou seja, {{percent_nao_classificavel}}%.

{% endif %}

* Como chegamos a esses valores limites: a acurácia do modelo foi obtida em amostra com {{num_amostra_apuracao}} configurações de base de teste por validação cruzada com repetições aleatórias, alcançando {{percent_acuracia_teste}}. Considerando-se a tabela A-6 do livro Estatística Experimental (Natrella, 1963), para uma probabilidade de 99% que 99,9% da população esteja na faixa, para o tamanho amostral de {{num_amostra_apuracao_assumida_definicao_limite_acc}}, encontramos esses limites como aceitáveis.